Essays on market design and regulation in electricity systems

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

2015

vorgelegt
von

Diplom-Ökonom Joachim Bertsch

aus

Esslingen am Neckar

Referent:	Prof. Dr. Felix Höffler
Koreferent:	Prof. Dr.-Ing. Christian Rehtanz
Tag der Promotion:	01.06.2016

To my big brother.

Acknowledgements

First of all, I want to thank Prof. Dr. Felix Höffler for supervising my thesis. His constructive support and feedback have been very important for my research over the last years. I also would like to thank Prof. Dr.-Ing. Christian Rehtanz for co-referring my thesis. Furthermore, my gratitude goes to Prof. Dr. Marc Oliver Bettzüge for chairing the examination committee. I thank PD Dr. Dietmar Lindenberger for substituting Prof. Dr. Höffler during the disputation.

I'm indebted to the Institute of Energy Economics at the University of Cologne, and therein especially to Prof. Dr. Marc Oliver Bettzüge for providing financial support and a very good research environment. Furthermore, I want to thank the administration, communication and IT department at the institute and the chairs, especially Clara Dewes, for their very helpful support.

Financial support of the ’Deutsche Forschungsgesellschaft’ (German research society) through research grant HO 5108/2-1 is gratefully acknowledged. Chapters 2 and 3 were funded by the German Federal Ministry for Economic Affairs and Energy based on ruling of the Deutsche Bundestag. This financial support through grant 03ESP239 is gratefully acknowledged. Despite the financial support, the responsibility for the content of the publication lays solely with the author(s).

I'm grateful for my colleagues at the institute, who are excellent both on a personal and professional level, and were essential for me to write and finish my thesis. My special thanks go to my (repeated) co-author Simeon Hagspiel, with whom researching was – and hopefully will be– always enjoyable and insightful. I also would like to thank Lisa Just for the joint work, especially her contributions in the field of positive economics. I thank Tom Brown for the great collaboration within our research projects. Discussions and Comments on my research by Christian Tode are highly appreciated.

Finally, I would like to thank my family, especially my parents for their support and for enabling me to write this thesis.

Joachim Bertsch June 2016

Contents

Introduction

Electricity is bound to a grid infrastructure connecting producers and consumers at different locations. While it is generally comprehensive that a single grid infrastructure suffices, there is also a common understanding in economics that in grid-bound industries the grid infrastructure should be provided by one single firm. As a consequence, regulation of the firm is required in order to prevent her from exploiting this (natural) monopoly status. In contrast, the production of electricity is supposed to be most efficient if many different, competing firms exist. Recognizing this, the tasks of electricity generation and transmission have been separated, leading to grid firms, which are regulated, and production firms competing in electricity markets. This so-called unbundling entails complicating issues of organizing an electricity system, because even though production and transmission are separated, they are still interdependent by means of transmitting electricity from the producer to the consumer. Key to a successful organization is the handling of information – on one hand in an adequate *market design*, where the exchange of information must reflect the interdependence of the grid and production firms to achieve efficient coordination, and on the other hand in the *regulation* of the monopolistic grid firm, where information is crucial for designing an adequate regulatory contract.

An adequate *market design* in order to coordinate the regulated grid firm and the competing generation firms should reveal information about the scarcities of transmission or production capacities as well as high demand in the system. Prices contain this information, because they are the result of transactions between market participants, wherein they reveal their respective value of electricity. Hence, locations where the price for electricity is high indicate a high value of electricity and consecutively, a high value for additional grid or production capacities. In turn, high prices provide incentives for the market participants to take action. Prices, which reflect all available information, lead to an efficient allocation of market activities. Mostly motivated by political reasons, current market designs do not necessarily consider this basic economic principle and thus, lead to distorted price signals and inefficient allocations. However, the distortion might prove useful in case the market fails to process available information. Either way, the price does not fully capture all information in the market resulting in a trade-off between the inefficiency induced by market design or by the market itself. Therefore, the best way information is reflected in the market design is ambiguous and depends on the specific circumstances.

Successful *regulation* depends on the distribution of information between the regulator and the grid firm where the latter one is usually better informed about how to operate and expand the infrastructure. A regulatory contract has to be designed such that the regulated firm is not able to exploit her informational advantage. However, since it is not possible to fully control and supervise the grid firm, occasionally some inefficiency has to be accepted. The task for the regulator is then to minimize these inefficiencies and provide incentives for the firm to behave socially optimal. Therein, the regulator faces the trade-off between granting the firm a (costly) rent and giving the firm less discretion. This in turn leads to misguided firm activities, e.g., excessive grid expansion, due to the lack of information of the regulator. The trade-off and thus, an efficient regulation is determined by the characteristics of the information – or more precisely – the information asymmetries.

This thesis studies the described issues of handling information by analyzing the market design – with a special focus on the information exchanged between the regulated grid firm and the generation firms – and the regulation of the grid firm. The thesis consists of four chapters, each based on a single paper to which the authors contributed equally:

1. Congestion management in power systems - Long-term modeling framework and large-scale application (with Simeon Hagspiel and Lisa Just). *EWI Working Paper 15/03, revised and resubmitted to the Journal of Regulatory Economics.*

2. The relevance of grid expansion under zonal markets (with Tom Brown, Simeon Hagspiel and Lisa Just). *EWI Working Paper 15/07, submitted to The Energy Journal.*

3. Is an inefficient transmission market better than none at all? - On zonal and nodal pricing in electricity systems. *EWI Working Paper 15/05, submitted to The Energy Journal.*

4. Regulation of non-marketed outputs and substitutable inputs (with Simeon Hagspiel). *EWI Working Paper 15/06, in preparation for submission to the International Journal of Industrial Organization.*

In the following, a short overview of the different chapters is given, before the methodology and the results are critically discussed and compared.

1.1 Outline

In **Chapter 2**, the market design regarding the exchange of information between competitive generation and regulated grid firms is analyzed. Different approaches exist, which handle the exchange in terms of congestion management and revealing scarcities in the grid in different ways. In a nodal pricing regime, there is one price per grid node, which contains all information on the scarcity of grid and production. While this approach is favored by the theoretical literature as the most efficient design, alternative approaches with a uniform price in a zone containing several grid nodes have been implemented, e.g., in Europe. Congestion within these zones is then handled by administrative measures requested from the grid firm. There are three main characteristics along which the different design can be described: First, the geographical scope of the price zone. Second, the way congestion in a zone is handled, i.e., by administrative measures (e.g. redispatch) or more market based approaches such as local grid tariffs for generation (generator- or g-component). Third, which information of grid capacities between the zones is made available to the market. This chapter analyzes the differences and performance of the market designs with different characteristics.

Our approach is two-fold: First, we develop a generalized and flexible economic modeling framework as a decomposed inter-temporal equilibrium model for operation and investment including generation, transmission and their inter-linkage. Second, we apply this model in a numerical simulation and provide a suitable solution algorithm for large-scale problems. In the modeling framework, different organization of Transmission System Operators (TSOs), market areas, redispatch, local grid payments for generators as well as different calculation methods for cross-border capacities are implemented. This allows to consistently analyze the long-term effects of different settings. The numerical simulation allows to apply this theoretical framework to a real-world example, namely the Central Western European (CWE) region, and to show the long-term effects and inefficiencies of different market designs, amongst others the current European market design and the efficient benchmark: nodal pricing.

With the formulation of the economic modeling framework we show that the decomposition of the problem in a generation and transmission level, the various market designs can be easily implemented by modifying the exchange of information between the two levels. For instance, marginal transmission costs may be neglected within a price zone, leaving investment decisions without considering the impacts on the grid. In such an approach, nodal pricing is necessarily the first best benchmark, due to the unrestricted exchange of information between the two levels. Any modification of the problem leads to less efficient results. However, which restriction causes the most inefficiencies depends on the respective parameter setting and cannot be easily derived from the theoretical framework. We analyze this issue further by applying the economic modeling framework in a numerical solution. Again, the decomposition of the generation and transmission level proves helpful for solving the problem. The iterative solution algorithm we propose consists of updating information between the two levels until convergence is achieved. This also allows the coupling with other models, such as the used Alternating Current (AC) grid model for calculating the electricity flows in the transmission grid. Applying the model to the CWE region, we analyze the inefficiencies that arise if the market design is not first best. We consider 70 nodes and 174 power lines and simulate the development of the electricity system up to the year 2030. Inefficiencies amount to up to 4.6% of total system costs until 2030 compared to the efficient benchmark. The main source of inefficiencies is the coordination of the TSOs' activities, i.e., whether they jointly or nationally optimize their grid, and low trading capacities offered to the market caused by the implemented market coupling. Hence, considering the efforts for a common European electricity market, policies should foster the cooperation of the TSOs to improve efficiency.

Chapter 3 adds to the analysis of market designs of Chapter 2 by further analyzing the particular market design of zonal pricing, which is currently applied in Europe. The focus lays on the effects on the European electricity system and the achieving of the European energy targets until 2030, if grid expansion is not or only partially possible. The European electricity market design consists of bidding zones, usually aligned to national borders, with a uniform price. The information contained in this uniform price is correct if there is no congestion within the zone – an assumption, however, that is often not valid. If there is congestion, the price signal is distorted and leads to inefficient market results. Curative measures (e.g., redispatch) relieve congestion and the inefficiencies in the short term, while grid expansion shall relieve congestion in the long term. However, if this grid expansion is delayed or restricted due to, e.g., public acceptance issues, the zonal market design fails to provide efficient long-term signals for scarce transmission capacities. A blueprint of this development is the case of Germany, where generation capacity is scarce at some locations due to the market design. Thus, it needs to be explicitly contracted forward by the TSO outside of the market to relieve local congestion, because the market participants only see an uniform price as a distorted price signal.

To quantify these effects, we expand the model developed in Chapter 2 to incorporate different levels of grid restrictions. The inter-temporal equilibrium model for operation and investment including generation, transmission and their inter-linkage represents the current European market design, i.e., zonal pricing with redispatch. We gradually restrict the allowed levels of grid expansion per decade in Europe in different scenarios to get insights into the functionality of zonal markets with restricted grid expansion.

We find severe effects if grid expansion is highly restricted. With no grid expansion at all, 2% - 3% of load has to be curtailed. Even with a significantly higher allowed grid expansion, the load cannot be served completely at all times. Most of the load curtailment takes place in Southern Germany. Curtailed load indicates missing capacity at some places, i.e., additional capacity would have to be contracted to prevent load curtailment (as done, e.g., in Germany through the so-called grid reserve). Restricted grid expansion also jeopardizes the EU 2030 climate targets due to a partial curtailment of renewables. For instance, in the scenario with no grid expansion the renewables share is 1.5 percentage points lower compared to any other scenario. In case of restricted overall capacity for grid expansion, Direct Current (DC) lines prove to be valuable due to possible point-to-point interconnections without extensive grid reinforcement elsewhere. The results show the shortfalls of zonal market design if grid expansion is restricted, which seems to be very likely in reality. Measures to prevent load curtailment require either administrative intervention by contracting generation capacity outside the market, a redefinition of zones or an implementation of locational price elements.

Chapter 4 proceeds with the analysis of market designs with different information handling, but – in contrast to the previous chapters –, studies an environment wherein the market can only partially process the available information. In efficient markets with full availability of information, one price per grid node is most efficient. However, empirical work has shown that such nodal pricing lacks efficiency if forward transmission markets come into play. Those markets tend to be not fully efficient due to, for instance, transaction costs or missing liquidity. In such situations, zonal market designs, which have inherently distorted uniform prices might be advantageous if the distortion counterbalances the induced inefficiencies. The trade-off between the two market designs is discussed in this chapter.

First, a simple two node model to analyze the general trade-off is developed. The model consists of two nodes and one transmission line including a spot and an energy forward market. The outcomes for nodal and zonal pricing are compared with respect to overall welfare. Next, the model is expanded by incorporating more nodes, loop flows and an additional transmission forward market. This stochastic equilibrium model allows to consistently analyze nodal pricing and zonal pricing with redispatch. The market inefficiency on the forward market is introduced via a bid-ask-spread and risk aversion of the market participants. Due to the analytical complexity, the model is solved numerically and the welfare impacts of a broad variation of relevant parameters are analyzed. These parameters include supply, demand and grid parameters as well as the parameters for the bid-ask-spread and risk aversion.

The analysis of efficient spot and forward markets confirms the results of the theoretical literature on nodal pricing: It is the efficient benchmark and any other design can only be as good as nodal pricing but never outperform it. In the analysis, this result is due to the inherent inefficiency induced by a producer-based redispatch in the zonal pricing regime. If consumers are allowed for redispatch, the outcome of both regimes is the same. These findings hold for the simple two node example as well as the more complex stochastic equilibrium model. Considering an inefficient transmission forward market for nodal pricing, i.e., by introducing a positive bid-ask-spread, there are parameter constellations for which zonal pricing shows a higher overall welfare than nodal pricing. Zonal pricing outperforms nodal pricing in situations where there is little congestion, and supply costs or demand elasticity are relatively low. The inefficiency of zonal pricing then causes only small welfare losses, because it comes at relatively low costs. Another result is that the inefficiency of transmission forward markets and reduced trading volume of forward transmission contracts also impacts the trading volumes of the energy forwards, despite these being traded at a central hub. These spillovers increase the negative welfare effect of inefficiencies in the transmission forward market. In conclusion, one should consider the possible inefficiency of nodal pricing regimes on the transmission forward market as a relevant criterion in the discussion on nodal and zonal pricing or the best definition of zones.

Chapter 5 studies the regulation of the grid firm in case the regulator and the regulated firm have different levels of information. Usually, the firm is better informed about the adequacy and efficiency of her actions, which leads to asymmetric information between the regulator and the firm. The provision of uninterrupted electricity transmission is a good example for this: While the output, i.e., a stable grid quality with no blackouts might be easily observable, the *efficient* level and combination of inputs, i.e., grid expansion or sophisticated grid operation, is difficult to be judged by the regulator. In this chapter, a theoretically optimal contract framework for the general case of such information asymmetries is developed and compared to approaches applied in practice, especially for electricity transmission.

We develop a theoretical principal-agent-model to determine the optimal regulatory strategy in terms of a Bayesian menu of contracts. The setting of the regulatory problem is such that one input and the output is observable, while the realization of the other input and its cost are not observable. The information asymmetries are represented by discrete distributions regarding the overall input levels as well as the marginal rate of substitution between the inputs. The analysis of this novel setting is compared to a simpler non-Bayesian approach often applied in practice. Thereby, we bridge the gap between the academic discussion and regulatory practice.

We find that the information asymmetries and shadow costs of public funding impede the implementation of the first best solution. The expected social welfare necessarily includes some additional rent for the firm. This rent is highest for the firm with an efficient production technology, for which input levels are first best in the optimal solution. For all other possible types, the observable input factor is distorted upwards in the optimal solution and hence, input mix and level deviate from the first best. For some parameter settings, it is optimal to offer the same contract for several types of the firm (bunching). A cost-based regulatory regime – as an example for a regulatory approach applied in practice for electricity transmission – may be close to the obtained second best solution if a high overall input level is very probable and shadow cost of public funding are large. This might indeed characterize the current situation so that the practical regulatory approach may be close to the theoretical optimum.

1.2 Discussion on methodology and results

Each chapter highlights a specific aspect of how information is handled in the market design and regulation in electricity systems. For each aspect a suitable methodology was chosen. Chapters 2 - 4 are based on a fundamental modeling approach, which is preceded by abstract economic formulations. The economic formulations in these chapters were chosen to show the general effects, which are then analyzed in more detail with a complex fundamental model. The fundamental approach is necessary due to the fact that the electricity grid plays a major role in the analyzed issues requiring a representation of the technical properties (loop flows etc.). In Chapter 5, the issue of regulation is analyzed with a highly stylized principal-agent-model, which was chosen due to the possibility to reflect the relevant information as well as to focus on the contractual relationship of the regulator and the regulated firm.

The approach of Chapters 2 and 3 rely on rather strong assumptions: market participants are assumed to behave fully rational and have perfect foresight. Furthermore, competitive, efficient markets and perfectly incentivized regulated firms are assumed. While some assumptions (e.g, competitiveness or perfect foresight) might not jeopardize the general conclusions, others (e.g., efficient markets) certainly do, as is shown in Chapter 4. Furthermore, the applied methodology in the numerical solution approach impacts the obtained results. The underlying problem is highly non-linear due to the technical properties of the electricity grid. It is not shown analytically, that the applied methodology converges to a unique, global equilibrium, although the numerical behavior suggests that this might be the case. In addition, the chosen convergence threshold is rather high regarding the obtained efficiency deltas of the different market designs. Related to this is the assumption of inelastic demand, necessary to achieve numerical tractability. While today this might be rather justified, it is doubtful whether this assumption will hold in the future. Certainly, all the general assumptions and the numerical solution approach impact the results and may alter the magnitude of the efficiency deltas and also the order obtained. Hence, the results and conclusions should be interpreted with care considering these shortfalls.

Some of the methodological shortfalls of Chapters 2 and 3 were picked up in Chapter 4, namely the assumptions of perfect foresight, efficient markets and elastic demand. Comparing the results of Chapters 2 and 3 to Chapter 4 shows that the assumptions are indeed critical. Introducing inefficient markets in Chapter 4, and thus, altering a central assumption of the previous chapters, makes the conclusion about the efficiency of the market design more ambiguous. Still, also this chapter keeps the assumptions of fully rational behavior, competitive markets as well as perfect regulation. Again, the solution approach to achieve a global equilibrium is not proven analytically, although the numerical results suggest again a robust equilibrium.

Chapter 5 focuses on the issue of information in a contract framework, where information about technical properties is less relevant than the distribution and asymmetry of the information. The applied theoretical principal-agent-model is able to capture several dimensions of information asymmetry. However, this comes at the costs of a high degree of abstraction. The Bayesian solution approach assumes common knowledge of the distributions and the level of information asymmetry. Although this is a necessary assumption for the model to solve it analytically, it is rather doubtful when it comes to an implementation of such regulatory contracts. This chapter elaborates on the incentives for the grid firm, which was assumed to be perfectly incentivized and behaving socially optimal in the former chapters. Looking at the insights from Chapter 5, namely the upwards distorted grid expansion from the efficient level, shows that the assumption of a perfectly incentivized grid firm is highly discussable. The results on grid expansion and operation would be most probably altered, if the impacts of the contract framework are included.

The discussion of the chosen methodology and the impacts on the results and conclusions shows that *general* conclusions require careful judgment about the relevance and effects of the underlying assumptions. This has to be especially taken into account if the results are transferred from the scientific sphere to a more practical application. Nevertheless, this thesis provides a consistent analysis of different aspects and thus, contributes to further advancements in the field of market design and regulation of electricity systems.

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

2.1 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 efficiency. This was shown in the seminal work of Schweppe et al. (1988) and Hogan (1992). Nevertheless, many markets deviate 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, the dispatch can be impacted by charging grid costs directly to generators in order to avoid congestion in the market clearing process (a so-called generator- or g-component, also known as grid connection charge). Such charges reflect the locational scarcity of the grid, and are thus conceptually similar to nodal prices, depending on the calculation method applied (see Brunekreeft et al. (2005) for a comprehensive discussion). 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 (Brunekreeft et al. (2005), Oggioni and Smeers (2012), Oggioni and Smeers (2013)).[1]

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 a single TSO or multi-lateral arrangements for counter-trading between several TSOs 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

[1] Under implicit market coupling, cross-border capacities and prices are implicitly taken into account during the joint clearing process of coupled markets.

Oren (2006) and Rious et al. (2009) formulate the coordination problem between a generation 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, though some inefficiencies remain and the first best is not realized. 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.[2] To tackle the complex nature of the optimization problem, we develop a numerical solution algorithm based on decomposition, while a detailed analysis of the convergence behavior suggests that the results obtained are robust. 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% 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.

[2] 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.

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.

2.2 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.[3] By design, this model does not contain any frictions and inefficiencies. Hence, the results obtained are 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).[4]

To depict various congestion management designs, we 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 then 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.[5] 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 or coupled zonal markets), the regulation and organization of TSOs (one single TSO for all zones or several zonal 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). We consider Net Transfer Capacity (NTC) and flow-based market coupling as cross-border capacity allocation algorithms

[3] 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.

[4] One main difference in our model is the assumption of an inelastic demand which was necessary to formulate and solve the model as a linear program. We will elaborate on this issue in Section 2.2.1.

[5] Efficient in this context means that the TSO(s) are perfectly regulated to expand and operate the grid at minimal costs.

because they have been used extensively in the European context (see, e.g., Glachant (2010)). 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. 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)), hence providing a much better consideration of the physical grid properties which is crucially important in case of meshed networks. As a consequence, more capacity can generally be offered for trading between markets, and a better usage of existing infrastructures is achieved. The analyzed settings are summarized in the following Table 2.1.

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

Table 2.1: Analyzed congestion management designs

Noticeably, despite the separated generation and transmission levels, agents are in all settings assumed to act rationally and simultaneously while taking into account the activities of the other stage.[6] 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 independent institution (e.g., the power exchange) being responsible for coordinating the activities of the different participating agents and for market clearing.[7] 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. We will thoroughly discuss issues of convexity in the context of the numerical implementation in Section 2.3. 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.[8]

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

[7] 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.

[8] In our numerical application, this approach is supplemented with discounted future cash flows. See Section 2.4 for further details.

For the sake of readability (and in contrast to the large-scale application presented in Section 2.4), we make some simplifications in the theoretical framework: dispatch decisions are realized in several points of time, but invest decisions are undertaken only once. Furthermore, we neglect different types of generation technologies that may be available at a node. This simplification does not change any of the conclusions drawn from the theoretical formulation.[9]

For developing the economic modeling framework in the following subsections, we will deploy parameters, variables and sets as depicted in Table 2.2 in the Appendix.

2.2.1 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 – leading to a welfare optimal electricity system. It hence represents the first best setting in our analysis of different congestion management designs. With the assumption of a social planner or perfect competition and regulation, nodal prices can be derived from locational marginal costs (of generation and capacity) in a market clearing that implicitly considers the physical properties of the electricity network (specifically, loop flows). Abstracting from economies of scale and lumpiness of investment, it can be shown that an efficient and unique equilibrium exists under nodal prices (Caramanis (1982), Joskow and Tirole (2005), Rious et al. (2009)). In line with these findings, we assume constant marginal grid costs as well as continuous generation and transmission expansion.[10] Another assumption in our formulation is an inelastic (yet time-varying) demand. The reason for assuming an inelastic demand is mainly triggered by the excessive computational burden that would be induced by an elastic demand in the numerical solution approach (an inelastic demand allows us to formulate and solve the model as a linear instead of a non-linear program). As a drawback, the assumption of an inelastic demand differs from the formulation in Schweppe et al. (1988) and leads to the artifact that demand can never set the price. However, scarcity rents to cover capacity costs are still possible under perfect information and competition (including entry and exit of generators). For instance, consider the bid of a peak load plant during a single peak load hour when it is dispatched and pivotal. The bid will consist of the variable costs plus the long-term marginal costs of the capacity. If the bid was lower, the peak load plant would leave the market due to an overall loss. If the bid was

[9] To include multiple instances in time for investments, the formulation could easily be adapted by adding an index to all parameters, variables and equations related to installed capacities (generation and transmission). In the same vein, an additional index could be inserted to account for different types of generation technologies.

[10] 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.

higher, another peak load plant would enter the market due to the possibility of making a profit. This forces the peak load plant to bid its true variable plus marginal capacity costs. Once accepted, this bid can be interpreted as the resulting market prices under capacity scarcity. Lastly, note that off-peak hours can also have capacity components in prices if there is a diversified mix of generation technologies, characterized by different cost structures.

The following optimization problem *P1* is similar to the formulation of an integrated problem for operating generation and transmission as in Schweppe et al. (1988), except for the major change of demand being inelastic. In this formulation, a social planner or an integrated firm minimizes total system costs of the operation and investment of generation and transmission.

P1 *Integrated Problem*

$$\min_{\overline{G}_i, G_{i,t}, T_{i,j,t}, \overline{P}_{i,j}} \quad X = \sum_i \delta_i \overline{G}_i + \sum_{i,t} \gamma_{i,t} G_{i,t} + \sum_{i,j} \mu_{i,j} \overline{P}_{i,j} \tag{2.1a}$$

$$\text{s.t.} \qquad G_{i,t} - \sum_j T_{i,j,t} = d_{i,t} \qquad \forall i,t \qquad | \, \lambda_{i,t} \tag{2.1b}$$

$$G_{i,t} \leq \overline{G}_i \qquad \forall i,t \tag{2.1c}$$

$$|T_{i,j,t}| = |P_{i,j,t}(\overline{P}_{k,l}, G_{k,t}, d_{k,t})| \leq \overline{P}_{i,j} \qquad \forall i,j,t \qquad | \, \kappa_{i,j,t} \tag{2.1d}$$

$$T_{i,j,t} = -T_{j,i,t} \qquad \forall i,j,t \tag{2.1e}$$

Indices i, j, k, l represent nodes in the system. Generation $G_{i,t}$, generation capacity $\overline{G}_i$, trade $T_{i,j,t}$ and transmission capacity $\overline{P}_{i,j}$ are optimization variables. 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 $\lambda_{i,t}$ of the equilibrium constraint which states that the demand level $d_{i,t}$ at node i can be either satisfied by generation at the same node or trade between nodes (Equation (2.1b)). Equations (2.1c) and (2.1d) mirror that generation is restricted by installed generation capacities, and physical flows by installed transmission capacities. Furthermore, trades from node i to node j are necessarily equal to negative trades from node j to node i (Equation (2.1e)). As the market clearing fully accounts for the transmission network in the nodal pricing regime, trade between adjacent nodes is equal to physical flows on the respective line, i.e., $T_{i,j,t} = P_{i,j,t}$ (Equation (2.1d)).

Load flows on transmission lines are based on Kirchhoff's law, which we represent based on a linearized load flow approach.[11] Thereby, flows are impacted by generation ($G_{k,t}$)

[11] We will use the PTDF approach shown in the Appendix in our numerical implementation in Section 2.3, as this enables a linearization of the generally non-linear load flow problem, given a fixed transmission network (cf. Hagspiel et al. (2014)).

and demand ($d_{k,t}$), i.e., power balances of all nodes in the system, as well as by the physical properties of the transmission system, represented by installed transmission capacities $\overline{P}_{k,l}$. Thus, there is a functional dependency of flows and trades on generation, demand, and line capacities throughout the system, i.e., $T_{i,j,t} = T_{i,j,t}(\overline{P}_{k,l}, G_{k,t}, d_{k,t})$.

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 and demand while respecting generation capacity constraints. As in *P1*, the same nodal prices are obtained by the dual variable $\lambda_{i,t}$ of the equilibrium constraint (2.2b). 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 $\kappa_{i,j,t}$ 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 (2.2g)). 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*). As trade is a function of $\overline{P}$, which in turn is the decision variable in the transmission problem, the market clearing conditions need to reoccur in the transmission problem.

P1'a *Generation*

$$\min_{\overline{G}_i, G_{i,t}, T_{i,j,t}} \quad X = \sum_i \delta_i \overline{G}_i + \sum_{i,t} \gamma_{i,t} G_{i,t} + \sum_{i,j,t} \kappa_{i,j,t} T_{i,j,t} \tag{2.2a}$$

$$\text{s.t.} \quad G_{i,t} - \sum_j T_{i,j,t} = d_{i,t} \qquad \forall i, t \qquad | \, \lambda_{i,t} \tag{2.2b}$$

$$G_{i,t} \leq \overline{G}_i \qquad \forall i, t \tag{2.2c}$$

$$T_{i,j,t} = -T_{j,i,t} \qquad \forall i, j, t \tag{2.2d}$$

P1'b *Transmission*

$$\min_{\overline{P}_{i,j}} \quad Y = \sum_{i,j} \mu_{i,j} \overline{P}_{i,j} \tag{2.2e}$$

$$\text{s.t.} \quad G_{i,t} - \sum_j T_{i,j,t} = d_{i,t} \qquad \forall i, t \tag{2.2f}$$

$$|T_{i,j,t}| = |P_{i,j,t}(\overline{P}_{k,l}, G_{k,t}, d_{k,t})| \leq \overline{P}_{i,j} \qquad \forall i, j, t \qquad | \, \kappa_{i,j,t} \tag{2.2g}$$

$$T_{i,j,t} = -T_{j,i,t} \qquad \forall i, j, t \tag{2.2h}$$

As can be seen, all terms of *P1* reappear in *P1'*, however, allocated to two separated levels. Mathematically, the equivalence of *P1* and *P1'* is shown in the Appendix, where the first order conditions of both formulations are compared.

2.2.2 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 the more sophisticated flow-based market coupling as cross-border capacity allocation algorithms (see Oggioni and Smeers (2013)). Under the latter regime, 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).[12] We assume here a perfectly discriminating 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 the originally dispatched generation capacity is still able to cover capital costs from the spot market result. 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.

[12] Redispatch is always feasible due to the fact that the TSO can foresee congestion and hence, counteract by expanding line capacities.

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 (2.3f), now takes place on zonal instead of nodal markets. The corresponding dual variable $\lambda_{m,t}$ represents zonal prices, which do not include any grid costs except for cross-border capacities. This is indicated by the term $\sum_{m,n,t} \kappa_{m,n,t} T_{m,n,t}$ instead of the nodal formulation (with $\kappa_{i,j,t}$) above. Transmission prices are determined on the transmission stage (Equation (2.3j)). 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.[13] In addition to grid expansion, the TSO may relieve intra-zonal congestion and optimize the situation by means of redispatch measures $R_{i,t}$ at costs of $\gamma_{i,t} R_{i,t}$.

P2a *Generation*

$$\min_{\overline{G}_i, G_{i,t}, T_{m,n,t}} \quad X = \sum_i \delta_i \overline{G}_i + \sum_{i,t} \gamma_{i,t} G_{i,t} + \sum_{m,n,t} \kappa_{m,n,t} T_{m,n,t} \tag{2.3a}$$

$$\text{s.t.} \quad \sum_{i \in \mathbf{I_m}} G_{i,t} - \sum_n T_{m,n,t} = \sum_{i \in \mathbf{I_m}} d_{i,t} \quad \forall m, t \quad | \, \lambda_{m,t} \tag{2.3b}$$

$$G_{i,t} \leq \overline{G}_i \quad \forall i, t \tag{2.3c}$$

$$T_{m,n,t} = -T_{n,m,t} \quad \forall m, n, t \tag{2.3d}$$

P2b *Transmission*

$$\min_{\overline{P}_{i,j}, R_{i,t}} \quad Y = \sum_{i,j} \mu_{i,j} \overline{P}_{i,j} + \sum_{i,t} \gamma_{i,t} R_{i,t} \tag{2.3e}$$

$$\text{s.t.} \sum_{i \in \mathbf{I_m}} G_{i,t} - \sum_n T_{m,n,t} = \sum_{i \in \mathbf{I_m}} d_{i,t} \quad \forall m, t \tag{2.3f}$$

$$|T_{i,j,t}| = |P_{i,j,t}(\overline{P}_{k,l}, G_{k,t}, R_{k,t}, d_{k,t})| \leq \overline{P}_{i,j} \quad \forall i, j, t \quad | \, \kappa_{i,j,t} \tag{2.3g}$$

$$\sum_{i \in \mathbf{I_m}} R_{i,t} = 0 \quad \forall m, t \tag{2.3h}$$

$$0 \leq G_{i,t} + R_{i,t} \leq \overline{G}_i \quad \forall i, t \tag{2.3i}$$

$$\kappa_{m,n,t} = g(\kappa_{i,j,t}) \tag{2.3j}$$

$$T_{m,n,t} = -T_{n,m,t} \quad \forall m, n, t \tag{2.3k}$$

[13] 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 ($\sum_{m,n,t} \kappa_{m,n,t} T_{m,n,t}$) would disappear and an additional constraint for trading would be implemented ($|T_{m,n,t}| \leq C_{m,n}, \forall m, n, t$). The restriction of trading volumes $C_{m,n,t}$ would be calculated on the transmission stage *P2b* via a constraint $C_{m,n} = h(\overline{P}_{i,j})$ instead of the prices $\kappa_{m,n,t}$. These prices would then be the dual variable of the volume constraint on the generation stage, and necessarily coincide with $\kappa_{m,n,t}$.

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 2.1, left hand side), *Setting I* and *II* are identical: There is only one element $i \in \mathbf{I_m}$, such that Equation 2.3h fixes variable R_i to zero. Equation 2.3i is then no longer relevant, and the cost term of redispatch in the objective function $(\sum_{i,t} \gamma_{i,t} R_{i,t})$ becomes zero. The only difference remaining between *1'b P2b* is then Equation 2.3j. However, due to $I = M$, it follows that $\kappa_{m,n,t} = \kappa_{i,j,t}$, which, inserted on the generation level, yields equivalence of problems *P1'* and problem *P2* for the chosen example.

Example for 3 nodes and 2 markets: Figure 2.1, right hand side, shows an electricity system consisting of two markets m and n, where m includes one node (1) and n two nodes $(2, 3)$ at a point in time t. Function g for calculating the transmission price $\kappa_{m,n,t}$ (Equation (2.3j)) between the markets has to be defined, e.g., by averaging the single line prices $\kappa_{m,n,t} = (\kappa_{1,2,t} + \kappa_{1,3,t})/2$. Still, the TSO cannot supply the locational fully differentiated prices $\kappa_{1,2,t}, \kappa_{1,3,t}$ and $\kappa_{2,3,t}$ 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.

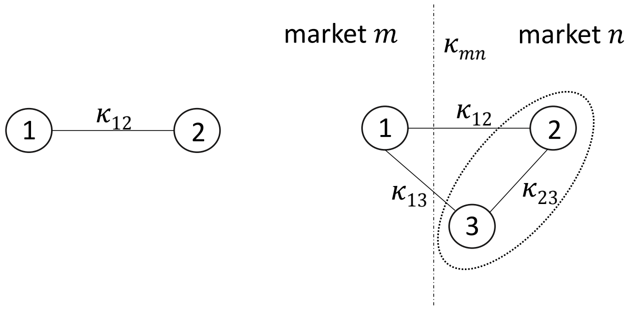


FIGURE 2.1: Two simple examples. Left: 2 nodes, 2 markets. Right: 3 nodes, 2 markets

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 $\kappa_{i,j,t}$ 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 $g(\kappa_{i,j,t})$. 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 2.4. 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 (2.3h)). 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.[14] Hence, the setting lacks locational signals for efficient generation investments within zonal markets.

2.2.3 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). Formally, problem *P3b*, now consists of multiple separate optimization problems for each zonal TSO, with the objective to minimize costs from zonal grid as well as from zonal redispatch measures. However, cross-border line capacities are also taken into account. As these 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. 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.

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

As a consequence, 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 or at most equal to those of *Setting II*.[15]

The mathematical program as well as further technical details of *Setting III* can be found in the Appendix.

2.2.4 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. A g-component charges grid costs directly to generators in order to avoid congestion in the market clearing process reflecting the impact of generators on the grid at each node and each instant of time. Thus, 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.

The mathematical program as well as further technical details of *Setting IV* can be found in the Appendix.

[15] The only mathematical difference of problem *P3b* compared to *P2b* is that the transmission level is partitioned into several optimization problems that are solved separately from each other. Hence, compared to problem *P2b* where the transmission level is solved comprehensively, this represents a more restrictive problem that must be inferior (or at most equal) to the one of *P3b*.

2.3 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 applied in the context of *Setting I* (Section 2.2.1), 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. Algorithmically, according to Benders (1962), 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 in the grid infrastructure have been made, according to Equation (2.15) and the PTDF calculation procedure presented in the Appendix. This iterative update of the grid properties, as applied in Hagspiel et al. (2014) and Ozdemir et al. (2015), 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.[16] In turn, linear problems can be solved effectively for global optima using standard techniques, such as the Simplex algorithm (e.g., Murty (1983)).

Even though the PTDF update ties in nicely with the iterative solution of the decomposed problem, it also imposes a particular challenge stemming from the non-linearity in the PTDF calculation (see Appendix). Specifically, despite the successive linearization and iterative solution, the non-linearity of the transmission expansion problem remains. Hence, neither the existence and uniqueness of a global optimum of the problem, nor the convergence of the solution algorithm can generally be guaranteed (e.g., Bazaraa et al. (2006)). This would change, however, if the problem was convex. Then, there would be a unique equilibrium, corresponding to a global optimum. Furthermore, deploying a Benders-type decomposition, the algorithm would preserve convexity and guarantee that the iterative solution converges towards this global optimum (Benders (1962) and,

[16] Accordingly, in our model $PTDF$ is depicted as a parameter that is updated in each iteration instead of a variable.

e.g., Conejo et al. (2006) for a general overview). Unfortunately, to the best of our knowledge, a formal proof of the (non-)convexity of the transmission expansion problem is still missing. Meanwhile, it would also be beyond the scope of this paper to approach this challenging problem. As an alternative, we build on numerical experience that has been gained by two papers that are closely related to ours in terms of the algorithmic approach: The analysis in Hagspiel et al. (2014) is closest to our application as they deploy the same successive PTDF update to co-optimize generation and transmission assets (including operation and investment). They show that the algorithm converges in a large number of configurations, including small analytically tractable test systems as well as large-scale applications. Furthermore, they do not detect issues of multiple equilibria in their analysis. In a very similar vein, Ozdemir et al. (2015) develop a methodology based on successive linear programming and Gauss-Seidel iteration to jointly optimize transmission and generation capacities. They report that even though they cannot guarantee convergence or global optimality either, their approach shows good performance. In the course of preparing the results presented in this paper, we were able to confirm the above findings in several model runs where we varied starting values over a broad range and did not find evidence neither against convergence nor against uniqueness of our optimum. Hence, even though not guaranteed, empirical evidence indicates that we are facing a numerical problem that we are able to reliably solve with our algorithm while converging towards an optimal solution. In our application, the obtained solution represents an intertemporal equilibrium without uncertainty. Interestingly, 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 follow the Benders decomposition approach described in Conejo et al. (2006), while considering the transmission capacities as complicating variables. We define the generation stage as the master problem, whereas the subproblem covers the transmission stage.[17] 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

[17] 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.

to the information they contain and hence determine to which 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 for investments, 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 benders cut (lower bound constraint) is added in each iteration u up to the current iteration v (Equation (2.4e)). This benders cut consists of total grid costs $Y^{(u)}$ as well as the marginal costs each unit of trade $T_{i,j,t}$ is causing in the grid per node, denoted by $\kappa^{(u)}_{i,j,t}$. 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

$$\min_{\overline{G}_i, G_{i,t}, T_{i,j,t}, \alpha} \quad X = \sum_i \delta_i \overline{G}_i + \sum_{i,t} \gamma_{i,t} G_{i,t} + \alpha \tag{2.4a}$$

$$\text{s.t.} \qquad G_{i,t} - \sum_j T_{i,j,t} = d_{i,t} \qquad \forall i, t \tag{2.4b}$$

$$G_{i,t} \leq \overline{G}_i \qquad \forall i, t \tag{2.4c}$$

$$T_{i,j,t} = -T_{j,i,t} \qquad \forall i, j, t \tag{2.4d}$$

$$Y^{(u)} + \sum_{i,j} \kappa_{i,j,t}^{(u)} \cdot (T_{i,j,t} - T_{i,j,t}^{(u)}) \leq \alpha \qquad \forall u = 1, ..., v-1 | v > 1 \tag{2.4e}$$

$$- -$$

$$G_{i,t}^{(v)} = \text{Optimal value of } G_{i,t} \qquad \forall i, t \tag{2.4f}$$

Sub-problem: transmission

$$\min_{\overline{P}_{i,j}, T_{i,j,t}} \quad Y = \sum_{i,j} \mu_{i,j} \overline{P}_{i,j} \tag{2.4g}$$

$$\text{s.t.} \quad |P_{i,j,t}| = \left| \sum_k PTDF_{k,i,j}^{(v)} \cdot (G_{k,t}^{(v)} - d_{k,t}) \right| \leq \overline{P}_{i,j} \quad \forall i, j, t \quad | \, \kappa_{i,j,t}^{(v)} \tag{2.4h}$$

$$- -$$

$$Y^{(v)} = \text{Optimal value of } Y \tag{2.4i}$$

$$PTDF^{(v)} = \text{PTDF matrix calculated based on } \overline{P}_{i,j} \tag{2.4j}$$

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 2.2 and is thus not discussed again in detail here.[18]

[18] 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 the Appendix.

2.4 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)). 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.

2.4.1 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 2.2 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% reduction with respect to 1990 levels in 2020, and 40% 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. Details on how the PTDF matrices are calculated can be found in the Appendix.

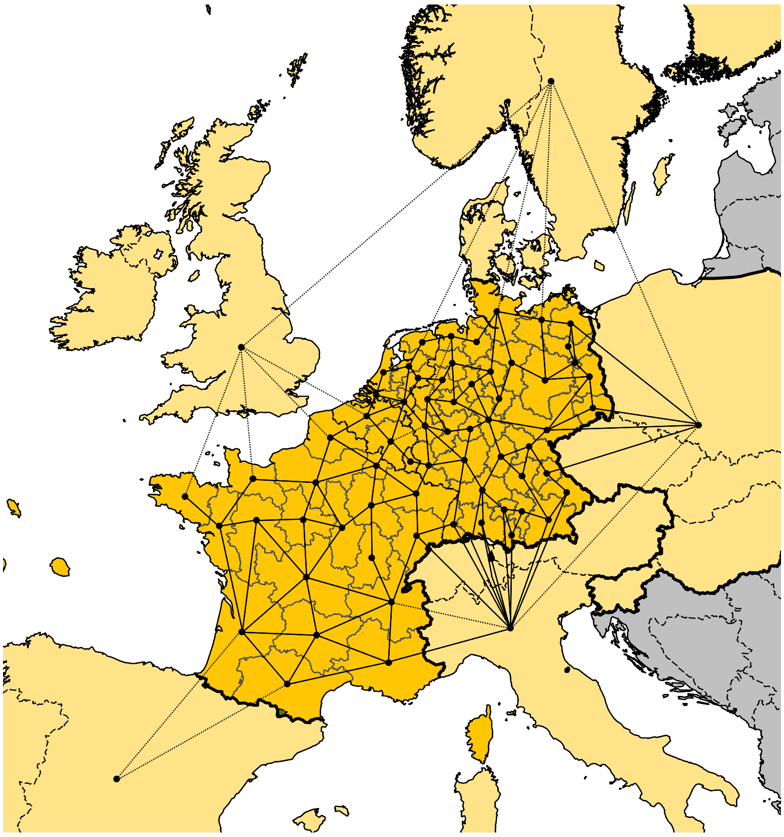


FIGURE 2.2: Representation of the CWE and neighboring regions in the model

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.[19] 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 the Appendix.

As in *Settings II-IV* zonal markets are being considered, assumptions about the cross-border price function $g(\kappa_{i,j,t})$ are necessary. For the NTC-based coupling of market

[19] 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.

zones, we define function $g(\kappa_{i,j,t}) = 1.43 \cdot \frac{\kappa_{i,j,t}\overline{P}_{i,j}}{\sum_{i,j}\overline{P}_{i,j}}$ $\forall i,j \in \mathbf{I_{m,cb}}$ 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.[20] 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.[21] The costs of cross-border lines are shared half-half by the two TSOs concerned, i.e., $\sigma_{i,j} = 0.5$.

2.4.2 Results and discussion

As usual in a Benders decomposition, we trace convergence based on the difference between an upper (i.e., the objective value of the integrated problem with solution values of the current iteration) and a lower bound (i.e., the objective of the master problem with the same solution values). We found that all settings undershoot a convergence threshold of 2.5% within 20 to 60 iterations (corresponding to a solution time of 2 to 7 days).[22] For practical reasons, we let all settings solve for one week and – after having double-checked that the convergence threshold of 2.5 % is met – take the last iteration to obtain our final results. The convergence threshold is chosen to keep the solution process computationally treatable, but is also based on empirical observations as well as expected convergence behavior. In fact, a lower convergence criterion increases computational time significantly, while further improvements on the objective value and optimized capacities are hardly observable.

To illustrate the convergent behavior of our problem, Figure 2.3, 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

[20] 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).

[21] Equation (2.24m) in the Appendix. 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).

[22] All models were coded in GAMS 24.2.2 and solved with CPLEX 12.6 on a High Performance Computer with two processors (1600 and 2700Mhz) and physical/virtual memory of 98/150GB.

observed to change only slightly, reaching change rates smaller than 0.01% after some 40 iterations. Moreover, as can be derived from the interpolation curves presented in Figure 2.3, left hand side, 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. Based on the fact that in a Benders decomposition the lower bound is non-decreasing (i.e., change rates are always positive as demonstrated in Figure 2.3, left hand side), and the empirically observed behavior of the lower bound, it can be concluded that the error further decreases mainly due to changes in the upper bound. Hence, we argue that the lower bound can be taken as a good approximation of the optimal objective value as soon as our convergence criterion is met. To support this argument and to deepen our insights, we closely analyzed optimized levels of the variables, observing that they reach fairly stable levels in the last iterations before reaching the convergence criterion.[23] As an example, the right hand side of Figure 2.3 shows aggregated AC line capacities obtained in the final runs of the nodal pricing setting.

Based on the interpolation curves estimated from the observed changes in the optimality error, a 1% threshold is expected to be reached after around 150 iterations. The estimated increase of the lower bound and hence, the improvement of the optimal solution, will then be around 0.21% higher compared to our obtained value. At around 300 iterations, the optimal solution will deviate by about 0.24% from our obtained value, and further improvements of the optimal solution would be negligible. Considering the extensive computational burden as well as the expected limited improvements, we do not consider a smaller convergence threshold and rather accept some level of uncertainty regarding the different levels of optimality achieved in the different settings.

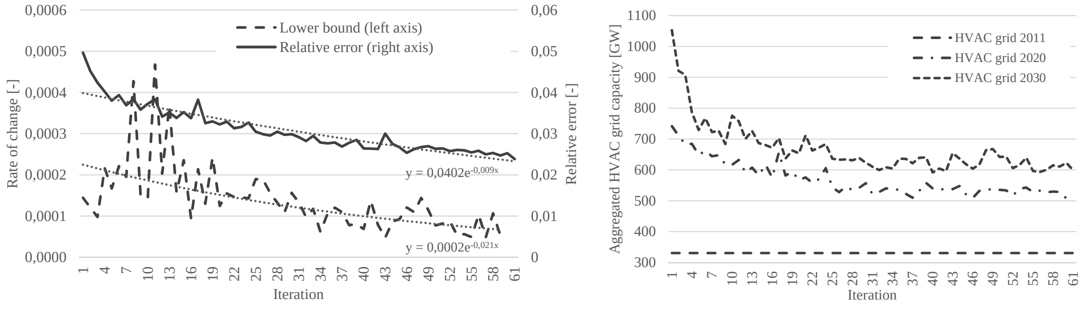


FIGURE 2.3: Development of lower bound, optimality error and aggregated AC line capacities during the iteration in *Setting I*

Costs are reported as accumulated discounted system costs.[24] 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

[23] Note that this argument is also supported by the analysis of convergence in a very similar setting published in Hagspiel et al. (2014).

[24] The discount rate is assumed to be 10% throughout all calculations.

are considered. Overall costs of electricity supply can be considered as a measure of efficiency and are reported in the following Figure 2.4 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.

Considering the optimality error in the obtained solution, it should be stressed that the exact differences reported here do not necessarily persist after full convergence. However, based on the above discussion about convergence, the general conclusions and order of magnitude are expected to remain valid.

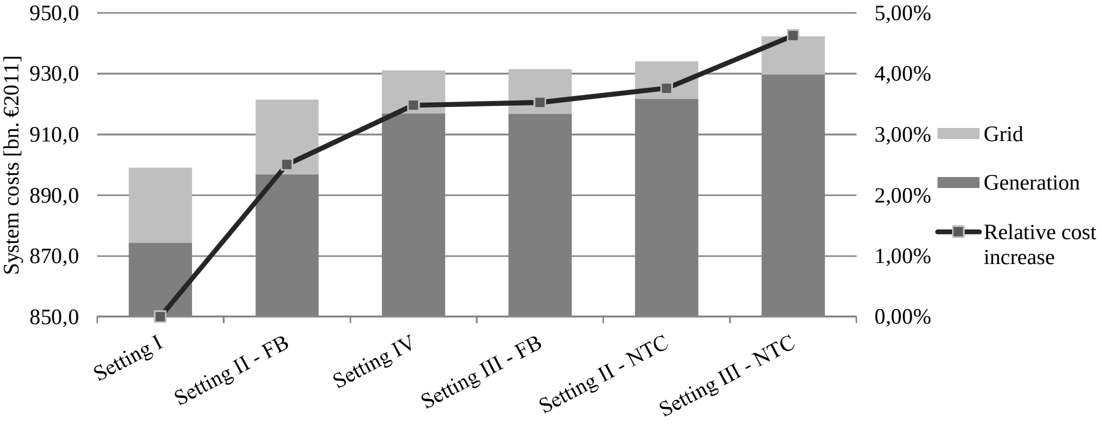


Figure 2.4: Total costs and relative performance of the different settings

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% relative to *Setting I* for the other settings. Thereby, NTC-based market coupling induces highest inefficiencies of 3.8% and 4.6% 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*).[25] 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% for the single TSO, respectively 3.5% for zonal TSOs compared to nodal pricing (*Setting II-FB* and *Setting III-FB*). Hence, efficiency gains of 1.1-1.3 % 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

[25] Since topology control (as, e.g., in Kunz (2013) is not considered, costs of redispatch could possibly be lower. However, since topology control would also be available in the market clearing of the nodal pricing, efficiency gains would persist for all regimes. Hence, the reported differences between the inefficiencies should be similar.

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% 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%)[26], the amount of grid capacities varies greatly between the different settings. Figure 2.5 shows the aggregated high voltage (HV) AC and HVDC line capacities.

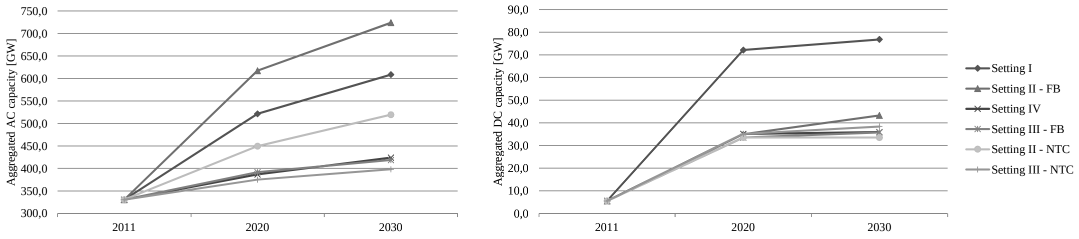


Figure 2.5: Aggregated line capacities AC and DC

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 331GW in 2011 to 398GW (*Setting III-NTC*) respectively 418GW (*Setting III-FB*) in 2030, corresponding to an increase of 20-28%. In case of a single TSO, cross-border along with overall line expansions are significantly higher compared to zonal TSOs, with 2030 levels reaching 519GW (*Setting II-NTC*) to 724GW (*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

[26] 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).

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 2.6, 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.[27]

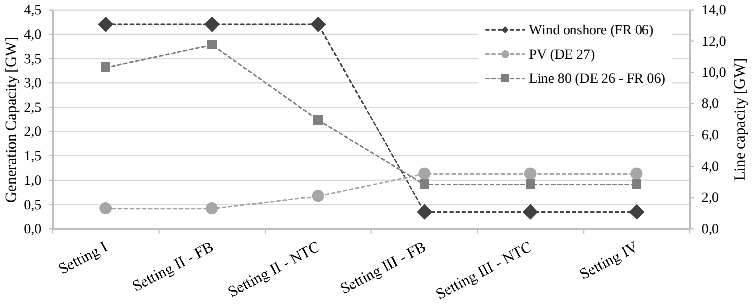


FIGURE 2.6: Exemplary grid expansion and regional allocation of renewable energies

2.5 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

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

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%. 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%. Reported cost differences might be impacted by numerical imprecision in the solution algorithm, although empirical observations of the convergence behavior suggest that the general effects as well as the order of magnitude persist. Noticeably, the magnitude 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.

2.6 Appendix

Notation list

Abbreviation	Dimension	Description
Model sets		
$i \in \mathbf{I}, j \in \mathbf{J}, k \in \mathbf{K}, l \in \mathbf{L}$		Nodes, $\mathbf{I}, \mathbf{J}, \mathbf{K}, \mathbf{L} = [1, 2, ...]$
$m, n \in \mathbf{M}$		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, $\mathbf{I_{m,cb}} \subset \mathbf{I_m}$, $\mathbf{J_{m,cb}} \subset \mathbf{J_m}$
$t \in \mathbf{T}$		Point in time for dispatch decisions (e.g., hours)
Model parameters		
δ_i	EUR/kW	Investment and FOM costs of generation capacity in node i
$\gamma_{i,t}$	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
$d_{i,t}$	kW	Electricity demand in node i
$PTDF_{k,i,j}$	$-$	Power Transfer Distribution Factor (impact of the power balance in node k on flows on line i, j)
$\sigma_{i,j}$	%	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, $\overline{G}_i \geq 0$
$G_{i,t}$	kW	Generation dispatch in node i, $G_{i,t} \geq 0$
$T_{i,j,t}, T_{m,n,t}$	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, $\overline{P}_{i,j} \geq 0 \;\; \forall \; i, j \neq j, i$
$P_{i,j,t}$	kW	Electricity flow on line between node i and node j
$R_{i,t}$	kW	Redispatch in node i
α	EUR	Helping variable to include transmission costs of the current iteration in the master problem
Model dual variables		
$\kappa_{i,j,t}, \kappa_{m,n,t}$	EUR/kW	price for transmission between nodes (i and j) or zones (m and n)
$\lambda_{i,t}, \lambda_{m,t}$	EUR/kW	nodal or zonal price for electricity

Table 2.2: Model sets, parameters and variables

Derivation of the load flow equations by means of PTDFs

Power Transfer Distribution Factors (PTDFs) are a well-established method to account for load flows in meshed electricity networks by means of linearization. They can be derived from the network equations in an AC power network that write as follows:[28]

$$
\begin{aligned}
P_i &= U_i \sum_{j \in \mathcal{I}} U_j (g_{i,j} \cos(\varphi_i - \varphi_j) + b_{i,j} \sin(\varphi_i - \varphi_j)) && (2.5)\\
Q_i &= U_i \sum_{j \in \mathcal{I}} U_j (g_{i,j} \sin(\varphi_i - \varphi_j) - b_{i,j} \cos(\varphi_i - \varphi_j)) && (2.6)\\
P_{i,j} &= U_i^2 g_{i,j} - U_i U_j g_{i,j} \cos(\varphi_i - \varphi_j) - U_i U_j b_{i,j} \sin(\varphi_i - \varphi_j) && (2.7)\\
Q_{i,j} &= \; - U_i^2 (b_{i,j} + b_{i,j}^{sh}) + U_i U_j b_{i,j} \cos(\varphi_i - \varphi_j) - U_i U_j g_{i,j} \sin(\varphi_i - \varphi_j). && (2.8)
\end{aligned}
$$

P_i and Q_i represent the net active and reactive power infeed (i.e., nodal power balances), and $P_{i,j}$ and $Q_{i,j}$ the active and reactive power flows between node i and j. Voltage levels U and phase angles φ of the nodes as well as series conductances g and series susceptances b of the transmission lines determine active and reactive power flows in a highly nonlinear way.

In order to linearize the above equations, a number of assumptions are made:

- All voltages are set to 1 p.u.
- Voltage angles are all similar (and hence, $\sin(\varphi_i - \varphi_j) \approx \varphi_i - \varphi_j$).
- Reactive power is neglected (i.e., $Q_i = Q_{i,j} = 0$).
- Losses are neglected and line reactances are much larger than their resistance, such that $x \gg r \approx 0$.

Under these assumptions and using Kirchoff's power law, the network equations can be simplified to

$$
\begin{aligned}
P_{i,j} &\approx \frac{1}{x_{i,j}} (\varphi_i - \varphi_j) && (2.9)\\
P_i &\approx \sum_{j \in \Omega_i} \frac{1}{x_{i,j}} (\varphi_i - \varphi_j), && (2.10)
\end{aligned}
$$

with Ω_i representing the nodes adjacent to i. If there are multiple nodes and branches, this can be written in a more convenient matrix notation as $\tilde{\boldsymbol{P_i}} = \tilde{\boldsymbol{B}} \cdot \tilde{\boldsymbol{\Theta}}$, with $\tilde{\boldsymbol{P_i}}$ being

[28] The following is based on Andersson (2011), even though the general approach can be found in most electrical engineering textbooks.

the vector of net active nodal power balances P_i, $\tilde{\boldsymbol{\Theta}}$ the vector of phase angles, and $\tilde{\boldsymbol{B}}$ the nodal admittance matrix with the following entries:

$$\tilde{B}_{i,j} = -\frac{1}{x_{i,j}} \tag{2.11}$$

$$\tilde{B}_{i,i} = \sum_{j \in \Omega_i} \frac{1}{x_{i,j}}. \tag{2.12}$$

By deleting the row and column belonging to the reference node (thus assuming a zero reference angle at this node), the previously singular matrix $\tilde{\boldsymbol{B}}$ becomes $\boldsymbol{B}$, the vector of phase angles $\boldsymbol{\Theta}$, and the vector of net active nodal power balances $\boldsymbol{P_i}$. We can now solve for $\boldsymbol{\Theta}$ by matrix inversion:

$$\boldsymbol{\Theta} = \boldsymbol{B}^{-1} \cdot \boldsymbol{P_i}. \tag{2.13}$$

Defining $H_{ki} = 1/x_{i,j}$, $H_{kj} = -1/x_{i,j}$ and $H_{km} = 0$ for $m \neq i, j$ (with k running over the branches i, j), Equation (2.9) can be rewritten in matrix form as $\boldsymbol{P_{i,j}} = \boldsymbol{H} \cdot \boldsymbol{\Theta}$. Inserting $\boldsymbol{\Theta}$ from Equation (2.13) finally yields

$$\boldsymbol{P_{i,j}} = \boldsymbol{H} \cdot \boldsymbol{\Theta} = \boldsymbol{H} \cdot \boldsymbol{B}^{-1} \cdot \boldsymbol{P_i} = PTDF \cdot \boldsymbol{P_i} \tag{2.14}$$

The elements of $PTDF$ are the power transfer distribution factors that constitute a linear relationship between nodal power balances and load flows. Note that the size of the $PTDF$ matrix is determined by the size of the system, with the number of matrix lines corresponding to the number of transmission lines, and the number of matrix columns representing the number of nodes. The matrix entry $PTDF_{k,i,j}$ represents the impact of the power balance in node k on power flows on line between node i and j. Also note that $PTDF$ essentially depends (only) on the line impedances $x_{i,j}$ in the system that in turn depend primarily on the respective line capacities $\overline{P}_{i,j}$. Hence, as done, e.g., in Hogan et al. (2010), we apply the law of parallel circuits to adjust line reactances when altering transmission capacities, i.e.,

$$x_{i,j} = \frac{\overline{P}^0_{i,j}}{\overline{P}_{i,j}} x^0_{i,j}, \tag{2.15}$$

where $\{\overline{P}^0_{i,j}, x^0_{i,j}\}$ is a point of reference taken from the original configuration of the transmission network. Overall, this yields a functional dependency of power flows on nodal balances (determined by generation G_k and load d_k in all nodes) as well as line capacities $\overline{P}_{k,l}$ of all lines in the system, i.e., $P_{i,j} = P_{i,j}(\overline{P}_{k,l}, G_k, d_k)$.

Equivalence of Problem P1 and P1'

To show the equivalence of the optimal solution of *P1* and *P1'*, we compare the problems by means of their Karush-Kuhn-Tucker (KKT)conditions. If they are equal, the optimal solution has to be equal, too (e.g., Bazaraa et al. (2006)). For the derivations, note that trade is a function of line capacity, generation and demand, i.e., $T_{i,j,t} = T_{i,j,t}(\overline{P}_{k,l}, G_{k,t}, d_{k,t})$, and that $T_{i,j,t} = -T_{j,i,t}$. The following is the Lagrangian function belonging to Problem *P1*:

$$\begin{aligned} L(\overline{G}_i, G_{i,t}, T_{i,j,t}, \overline{P}_{i,j}, \lambda_{i,t}, \tau_{i,t}, \kappa_{i,j,t}) = & \sum_i \delta_i \overline{G}_i + \sum_{i,t} \gamma_{i,t} G_{i,t} + \sum_{i,j} \mu_{i,j} \overline{P}_{i,j} \\ & + \sum_{i,t} (\lambda_{i,t}(G_{i,t} - \sum_j T_{i,j,t} - d_{i,t}) + \tau_{i,t}(G_{i,t} - \overline{G}_i)) \\ & + \sum_{i,j,t} (\kappa_{i,j,t}(|T_{i,j,t}| - \overline{P}_{i,j})) \end{aligned} \tag{2.16}$$

The corresponding KKT conditions are:

$$\tfrac{\partial L}{\partial \overline{G}_i} = \delta_i - \sum_t \tau_{i,t} \leq 0, \ \overline{G}_i \geq 0, \ \overline{G}_i(\frac{\partial L}{\partial \overline{G}_i}) = 0 \quad \forall i \tag{2.17a}$$

$$\begin{aligned} \tfrac{\partial L}{\partial G_{i,t}} & = \gamma_{i,t} + \lambda_{i,t}(1 - \sum_j \frac{\partial T_{i,j,t}}{\partial G_{i,t}}) + \tau_{i,t} + \sum_j \kappa_{i,j,t} \frac{\partial T_{i,j,t}}{\partial G_{i,t}} \leq 0, \\ & G_{i,t} \geq 0, \ G_{i,t}(\frac{\partial L}{\partial G_{i,t}}) = 0 \ \forall i,t \end{aligned} \tag{2.17b}$$

$$\begin{aligned} \tfrac{\partial L}{\partial \overline{P}_{i,j}} & = \mu_{i,j} - \sum_t \lambda_{i,t} \frac{\partial T_{i,j,t}}{\partial \overline{P}_{i,j}} + \sum_t \kappa_{i,j,t}(\frac{\partial T_{i,j,t}}{\partial \overline{P}_{i,j}} - 1) \leq 0, \\ & \overline{P}_{i,j} \geq 0, \ \overline{P}_{i,j}(\frac{\partial L}{\partial \overline{P}_{i,j}}) = 0 \quad \forall i,j \end{aligned} \tag{2.17c}$$

$$\tfrac{\partial L}{\partial \kappa_{i,j,t}} = |T_{i,j,t}| - \overline{P}_{i,j} \leq 0, \quad \kappa_{i,j,t} \geq 0, \quad \kappa_{i,j,t}(\frac{\partial L}{\partial \kappa_{i,j,t}}) = 0 \qquad \forall i,j,t \tag{2.17d}$$

$$\tfrac{\partial L}{\partial \tau_{i,j}} = G_{i,t} - \overline{G}_i \leq 0, \quad \tau_{i,j} \geq 0, \quad \tau_{i,j}(\frac{\partial L}{\partial \tau_{i,j}}) = 0 \qquad \forall i,j \tag{2.17e}$$

$$\tfrac{\partial L}{\partial \lambda_{i,t}} = G_{i,t} - \sum_j T_{i,j,t} - d_{i,t} = 0 \qquad \forall i,t \tag{2.17f}$$

$$\tfrac{\partial L}{\partial T_{i,j,t}} = \kappa_{i,j,t} - \lambda_{i,t} + \lambda_{j,t} = 0 \qquad \forall i,j,t \tag{2.17g}$$

The Langragian functions for *P1'* are:

$$\begin{aligned} L'^a(\overline{G}_i, G_{i,t}, T_{i,j,t}, \lambda_{i,t}, \tau_{i,t}) = & \sum_i \delta_i \overline{G}_i + \sum_{i,t} \gamma_{i,t} G_{i,t} + \sum_{i,j,t} \kappa_{i,j,t} T_{i,j,t} \\ & + \sum_{i,t} (\lambda_{i,t}(G_{i,t} - \sum_{j,t} T_{i,j,t} - d_{i,t}) + \kappa_{i,j,t}(G_{i,t} - \overline{G}_i)) \end{aligned} \tag{2.18}$$

$$L'^b(\overline{P}_{i,j}, \kappa_{i,j,t}) = \sum_{i,j} \mu_{i,j}\overline{P}_{i,j} + \sum_{i,t}(\lambda_{i,t}(G_{i,t} - \sum_{j,t} T_{i,j,t} - d_{i,t})) + \sum_{i,j,t}(\kappa_{i,j,t}(|T_{i,j,t}| - \overline{P}_{i,j})) \tag{2.19}$$

The KKT conditions of *P1'a* are:

$$\frac{\partial L'^a}{\partial \overline{G}_i} = \delta_i - \sum_t \tau_{i,t} \leq 0, \ \overline{G}_i \geq 0, \ \overline{G}_i(\frac{\partial L}{\partial \overline{G}_i}) = 0 \quad \forall i \tag{2.20a}$$

$$\frac{\partial L'^a}{\partial G_{i,t}} = \gamma_{i,t} + \lambda_{i,t}(1 - \sum_j \frac{\partial T_{i,j,t}}{\partial G_{i,t}}) + \tau_{i,t} + \sum_j \kappa_{i,j,t}\frac{\partial T_{i,j,t}}{\partial G_{i,t}} \leq 0, \tag{2.20b}$$

$$G_{i,t} \geq 0, \ G_{i,t}(\frac{\partial L}{\partial G_{i,t}}) = 0 \ \forall i, t$$

$$\frac{\partial L'^a}{\partial \tau_{i,j}} = G_{i,t} - \overline{G}_i \leq 0, \quad \tau_{i,j} \geq 0, \quad \tau_{i,j}(\frac{\partial L'^a}{\partial \tau_{i,j}}) = 0 \qquad \forall i, j \tag{2.20c}$$

$$\frac{\partial L'^a}{\partial \lambda_{i,t}} = G_{i,t} - \sum_j T_{i,j,t} - d_{i,t} \qquad \forall i, t \tag{2.20d}$$

$$\frac{\partial L'^a}{\partial T_{i,j,t}} = \kappa_{i,j,t} - \lambda_{i,t} + \lambda_{j,t} = 0 \qquad \forall i, j, t \tag{2.20e}$$

The KKT conditions of *P1'b* are:

$$\frac{\partial L'^b}{\partial \overline{P}_{i,j}} = \mu_{i,j} - \sum_t \lambda_{i,t}\frac{\partial T_{i,j,t}}{\partial \overline{P}_{i,j}} + \sum_t \kappa_{i,j,t}(\frac{\partial T_{i,j,t}}{\partial \overline{P}_{i,j}} - 1) \leq 0, \tag{2.21a}$$

$$\overline{P}_{i,j} \geq 0, \ \overline{P}_{i,j}(\frac{\partial L}{\partial \overline{P}_{i,j}}) = 0 \quad \forall i, j$$

$$\frac{\partial L'^b}{\partial \kappa_{i,j,t}} = |T_{i,j,t}| - \overline{P}_{i,j} \leq 0, \quad \kappa_{i,j,t} \geq 0, \quad \kappa_{i,j,t}(\frac{\partial L}{\partial \kappa_{i,j,t}}) = 0 \qquad \forall i, j, t \tag{2.21b}$$

Comparing the KKT conditions of problem *P1* to the ones of *P1a* and *P1b*, we can conclude that the problems are indeed equivalent.

Model of Setting III: coupled zonal markets with zonal TSOs and zonal redispatch

Mathematically, the model of *Setting III*, representing coupled zonal markets with zonal TSOs and zonal redispatch, is formulated as follows:

P3a *Generation*

$$\min_{\overline{G}_i, G_{i,t}, T_{m,n,t}} \quad X = \sum_i \delta_i \overline{G}_i + \sum_{i,t} \gamma_{i,t} G_{i,t} + \sum_{m,n,t} \kappa_{m,n,t} T_{m,n,t} \tag{2.22a}$$

$$\text{s.t.} \quad \sum_{i \in \mathbf{I_m}} G_{i,t} - \sum_{n,t} T_{m,n,t} = \sum_{i \in \mathbf{I_m}} d_{i,t} \qquad \forall m, t \qquad | \, \lambda_m \tag{2.22b}$$

$$G_{i,t} \leq \overline{G}_i \qquad \forall i, t \tag{2.22c}$$

$$T_{m,n,t} = -T_{n,m,t} \qquad \forall m, n, t \tag{2.22d}$$

P3b *Transmission*

$$
\begin{aligned}
\min_{\overline{P}_{i,j\in\mathbf{I_m}},R_{i\in\mathbf{I_m},t}} \quad Y_m = \quad & \textstyle\sum_{i,j\in\mathbf{I_m}} \mu_{i,j}\overline{P}_{i,j} + \sum_{i,j\in\mathbf{I_{m,cb}}} \sigma_{i,j}\mu_{i,j}\overline{P}_{i,j} \\
& + \textstyle\sum_{i\in\mathbf{I_m},t} \gamma_{i,t}R_{i,t} \quad \forall m && (2.22e)\\
\text{s.t.} \quad & \textstyle\sum_{i\in\mathbf{I_m}} G_{i,t} - \sum_n T_{m,n,t} = \sum_{i\in\mathbf{I_m}} d_{i,t} \qquad \forall m,t && (2.22f)\\
|T_{i,j,t}| = \quad & |P_{i,j,t}(\overline{P}_{k,l}, G_{k,t}, R_{k,t}, d_{k,t})| \leq \overline{P}_{i,j} \; \forall t,i,j \in \mathbf{I_m} \;\; | \kappa_{i,j\in\mathbf{I_m}} && (2.22g)\\
& \textstyle\sum_{i\in\mathbf{I_m},t} R_{i,t} = 0 && (2.22h)\\
& 0 \leq G_{i,t} + R_{i,t} \leq \overline{G}_i \qquad \forall t, i \in \mathbf{I_m} && (2.22i)\\
& \kappa_{m,n,t} = g(\kappa_{i,j,t}) && (2.22j)\\
& T_{m,n,t} = -T_{n,m,t} \qquad \forall m,n,t && (2.22k)
\end{aligned}
$$

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 (2.22e)). For the redispatch, TSOs have to consider the same restrictions as in the previous setting (Equations (2.22h) and (2.22i)). TSOs are assumed to negotiate about the extension of cross-border capacities according to some 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. 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 (2.22j), prices for transmission between zones that are provided to the generation stage ($\kappa_{m,n,t}$) 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 $\kappa_{i,j,t}$ may now deviate from *Setting II* as they result from the separated activities of each zonal TSO (specifically, from Equation (2.22g), i.e., the restriction of flows on intra-zonal and cross-border lines).

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

Mathematically, the model of *Setting IV*, representing coupled zonal markets with zonal TSOs and g-component, is formulated as follows:

P4a *Generation*

$$\min_{\overline{G}_i, G_{i,t}, T_{m,n,t}} \quad X = \sum_i \delta_i \overline{G}_i + \sum_{i,t} \gamma_{i,t} G_{i,t} + \sum_{i,j,t} \kappa_{i,j,t} T_{i,j,t} \tag{2.23a}$$

$$\text{s.t.} \quad \sum_{i \in \mathbf{I_m}} G_{i,t} - \sum_n T_{m,n,t} = \sum_{i \in \mathbf{I_m}} d_{i,t} \qquad \forall m, t \qquad | \, \lambda_m \tag{2.23b}$$

$$G_{i,t} \leq \overline{G}_i \qquad \forall i, t \tag{2.23c}$$

$$T_{m,n,t} = -T_{n,m,t} \qquad \forall m, n, t \tag{2.23d}$$

P4b *Transmission*

$$\min_{\overline{P}_{i,j \in \mathbf{I_m}, \mathbf{I_{m,cb}}}} \quad Y_m = \sum_{i,j \in \mathbf{I_m}} \mu_{i,j} \overline{P}_{i,j} + \sum_{i,j \in \mathbf{I_{m,cb}}} \sigma_{i,j} \mu_{i,j} \overline{P}_{i,j} \qquad \forall m \tag{2.23e}$$

$$\text{s.t.} \sum_{i \in \mathbf{I_m}} G_{i,t} - \sum_n T_{m,n,t} = \sum_{i \in \mathbf{I_m}} d_{i,t} \qquad \forall m, t \tag{2.23f}$$

$$|T_{i,j,t}| = |P_{i,j,t}(\overline{P}_{k,l}, G_{k,t}, d_{k,t})| \leq \overline{P}_{i,j} \quad \forall t, i, j \in \mathbf{I_m}, \mathbf{I_{m,cb}} \quad | \, \kappa_{i,j \in \mathbf{I_m}, \mathbf{I_{m,cb}}, t} \tag{2.23g}$$

$$T_{m,n,t} = -T_{n,m,t} \qquad \forall m, n, t \tag{2.23h}$$

Problem *P4a* is almost identical to *P2a* (and *P3a*), with the exception of one term in the objective function (2.23a). With a g-component, generators pay nodal instead of zonal prices for transmission ($\kappa_{i,j,t}$ instead of $\kappa_{m,n,t}$), depending on the impact of their nodal generation level on the grid infrastructure (by means of $T_{i,j,t} = T_{i,j,t}(G_{k,t}, d_{k,t})$). These prices are determined by the zonal TSOs via their flow-restriction (2.23g).

Numerical algorithm for NTC-coupled zonal markets, zonal TSOs, and zonal redispatch

In the previous section, 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 $\kappa^{(v)}_{m,n,t}$ 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 $\kappa^{(v)}_{m,n,t}$ is calculated via a different function g_{FB}.

$v = 1$; convergence=false

While(convergence=false) {

Master problem: generation

$$\min_{\overline{G}_i, G_{i,t}, T_{m,n,t}, \alpha} \quad X = \sum_i \delta_i \overline{G}_i + \sum_{i,t} \gamma_{i,t} G_{i,t} + \alpha \tag{2.24a}$$

$$\text{s.t.} \quad \sum_{i \in \mathbf{I_m}} G_{i,t} - \sum_n T_{m,n,t} = \sum_{i \in \mathbf{I_m}} d_{i,t} \qquad \forall m, t \tag{2.24b}$$

$$G_{i,t} \leq \overline{G}_i \qquad \forall i, t \tag{2.24c}$$

$$T_{m,n,t} = -T_{n,m,t} \qquad \forall m, n, t \tag{2.24d}$$

$$\sum_m Y_m^{(u)} + \sum_{m,n,t} \kappa_{m,n,t}^{(u)} \cdot (T_{m,n,t} - T_{m,n,t}^{(u)}) \leq \alpha \qquad \forall u = 1, ..., v-1 | v > 1 \tag{2.24e}$$

$$- -$$

$$G_{i,t}^{(v)} = \text{Optimal value of } G_{i,t} \qquad \forall i, t \tag{2.24f}$$

Sub-problem: transmission

$$\min_{\overline{P}_{i,j \in \mathbf{I_m}, \mathbf{I_{m,cb}}}, R_{i \in \mathbf{I_m}, t}, T_{m,n,t}} \quad Y_m = \sum_{i,j \in \mathbf{I_m}} \mu_{i,j} \overline{P}_{i,j} + \frac{1}{2} \sum_{i,j \in \mathbf{I_{m,cb}}} \mu_{i,j} \overline{P}_{i,j} + \sum_{i \in \mathbf{I_m}, t} R_{i,t} \gamma_{i,t} \ \forall m \tag{2.24g}$$

$$\text{s.t.} \quad |P_{i,j,t}| = \left| \sum_k PTDF_{k,i,j} \cdot (G_{k,t}^{(v)} + R_{k,t} - d_{k,t}) \right| \leq \overline{P}_{i,j} \qquad \forall i, j, t \qquad | \ \kappa_{i,j,t}^{(v)} \tag{2.24h}$$

$$0 \leq R_{i,t} + G_{i,t}^{(v)} \leq \overline{G}_i \qquad \forall i, t \in \mathbf{I_m} \tag{2.24i}$$

$$\sum_{i \in \mathbf{I_m}} R_{i,t} = 0 \tag{2.24j}$$

$$- -$$

$$Y_m^* = \text{Optimal value of } Y_m \tag{2.24k}$$

$$PTDF^{(v)} = \text{New PTDF matrix calculated based on } \overline{P}_{i,j} \kappa_{m,n,t} = g_{NTC}(\kappa_{i,j,t}^{(v)}) \tag{2.24l}$$

$$\overline{P}_{i,j \in \mathbf{I_{m,cb}}} = \overline{P}_{i,j \in \mathbf{I_{n,cb}}} = \min \left\{ \overline{P}_{i,j \in \mathbf{I_{m,cb}}}; \overline{P}_{i,j \in \mathbf{I_{n,cb}}} \right\} \tag{2.24m}$$

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 2.3: 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	1253	1188
Wind Offshore (<20m depth)	2800	2350
Wind Offshore (>20m depth)	3080	2585
Photovoltaics (roof)	1260	935
Photovoltaics (ground)	1110	785
Biomass gas	2398	2395
Biomass solid	3297	3295
Biomass gas, CHP	2597	2595
Biomass solid, CHP	3497	3493
Geothermal	10504	9500
Compressed Air Storage	1100	1100
Pump Storage	1200	1200
Lignite	1500	1500
Lignite Innovative	1600	1600
Coal	1200	1200
Coal Innovative	2025	1800
IGCC	1700	1700
CCGT	711	711
OCGT	400	400
Nuclear	3157	3157

TABLE 2.4: 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 2.5: 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	1250 €/(MW*km)	6.3 €/(MW*km)
DC submarine	1100 €/(MW*km)	5.5 €/(MW*km)
DC converter pair	150000 €/MW	750.0 €/MW

TABLE 2.6: Assumptions for the grid extension and FOM costs

The relevance of grid expansion under zonal markets

3.1 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 the short term, this incompleteness is adressed by an administrative redispatch of generation facilities: After the market clearing, generators causing bottlenecks are requested to reduce their generation, while others increase their generation in order to achieve physical feasibility in the transmission 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.[29]

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, e.g., due to public opposition or extensive approval processes.[30] 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, even though the (zonal) market should ensure that sufficient capacity is installed to meet demand on a zonal level, the spatial allocation of generation units might not allow to transport electricity to the customers due to missing grid capacities. Thus, missing grid expansion might severly jeopardize the long-term

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

[30] See, e.g., the monitoring of the Ten Year Network Development Plan (ENTSO-E, 2015), where about 30 % of the projects are reported delayed or rescheduled, or Monitoringbericht (2013) for the case of Germany, where over 50 % of the projects are reported delayed.

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.[31] 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.[32] 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 3.1 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.

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]	-	-	4956	4604	5197	5253
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 3.1: Development of renewable share, redispatch measures and grid reserve in Germany[33]

The German case illustrates that the zonal market design may lead to severe effects, if grid expansion cannot be realized. Hence, in this paper, we analyze the relevance of grid expansion in the European zonal market design in the light of the EU's 2030 energy

[31] See (Burstedde, 2012) for a detailed discussion of the (in-)efficiencies of several redispatch designs.

[32] The European Union (EU) formulated an ambitious 2030 energy strategy, including a EU domestic reduction of greenhouse gases (GHG) by 40 % compared to 1990, a share of 27 % renewable energy, and a 27 % reduction in energy consumption compared to 2005 (European Commission, 2014a).

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

strategy. For this, we build on a long-term fundamental model of the European electricity market developed in Bertsch et al. (2015), allowing the representation of the European flow-based coupled zonal markets with redispatch. The model includes a generation dispatch, power flows, as well as generation and grid investments. Extending their approach, we implement restrictions for grid expansion per decade that are gradually tightened. Although, several authors analyzed *optimal* grid expansion in a European context (e.g., Schaber et al. (2012), Fürsch et al. (2013) or Hagspiel et al. (2014)), we are not aware of any literature analyzing the impacts of restricted grid expansion in the European zonal market design.

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, load cannot be served completely and load curtailment levels amounts up to 2 % (3 %) for 2020 (2030). Also with less restricted grid expansion, load curtailment is still above 0.2 % for scenarios 15 TWkm in 2020. In 2030, however, significant load curtailment only occurs for the scenarios of restrictions of up to 5 TWkm. The most load curtailment takes place in Southern Germany. Thereby, curtailment indicates 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 % of photovoltaic (PV) generation and over 3 % 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 prevent the expensive and politically unwanted curtailment of load, e.g., by contracting capacity outside the market. Alternatively, a redefinition of zones or introduction of locational price elements may be a suitable way to effectively reduce the amount of administrative intervention.

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

3.2 Methodology

3.2.1 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. (2015) and combine a cost-minimizing dynamic linear investment and dispatch market model with a model of the AC grid using a linear 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 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. In the second stage, transmission capacities are expanded and generation is redispatched 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 (3.1a)-(3.1l) state a simplified yet representative formulation of the problem:[34,35] 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 $i \in I_m$.[36] 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 $\overline{G}_{i,y}$.

[34] 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).

[35] A detailed overview containing all parameters, variables and sets is depicted in Table 3.5 in the Appendix.

[36] 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.

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 3.1c). 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 3.1d).

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 3.1g). The sum of all (positive and negative) redispatch measures has to amount to zero (Equation 3.1h) 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 by transmission capacities $\overline{P}_{m,n,y}$ (Equation 3.1i).[37] The exchange between the generation and the transmission stage takes place via the inter-zonal transmission capacities $\overline{P}_{m,n,t}$.[38] 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.[39] Due to the functional relationship of trades and transmission capacities, the market clearing condition has to reoccur on the transmission stage (Equation 3.1f). Trades from zone m to n are again equal to the negative trade from zone n to m (Equation 3.1l).

Generation

$$\min_{\overline{G}_{i,y}, G_{i,t}, T_{m,n,t}} \quad X = \sum_{i,y} \delta_{i,y} \overline{G}_{i,y} + \sum_{i,t} \gamma_{i,t} G_{i,t} \tag{3.1a}$$

$$\text{s.t.} \quad \sum_{i \in \mathbf{I_m}} G_{i,t} - \sum_{n} T_{m,n,t} = \sum_{i \in \mathbf{I_m}} d_{i,t} \qquad \forall m, t \tag{3.1b}$$

$$G_{i,t} \leq \overline{G}_{i,y} \qquad \forall i, t \tag{3.1c}$$

$$T_{m,n,t} = -T_{n,m,t} \leq \overline{P}_{m,n,t} \qquad \forall m, n, t \tag{3.1d}$$

[37] In our case power flows are represented by PTDFs that are treated as a parameter while solving the transmission stage, such that Equation (3.1i) 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.

[38] Note that this approach differs from Bertsch et al. (2015), where the exchange worked via transmission capacity marginals.

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

Transmission

$$\min_{\overline{P}_{i,j,y},R_{i,t}} \quad Y = \sum_{i,j,y} \mu_{i,j,y}\overline{P}_{i,j,y} + \sum_{i,t} \gamma_{i,t}R_{i,t} \tag{3.1e}$$

$$\text{s.t.} \quad \sum_{i\in\mathbf{I_m}} G_{i,t} - \sum_{n} T_{m,n,t} = \sum_{i\in\mathbf{I_m}} d_{i,t} \qquad \forall m,t \tag{3.1f}$$

$$0 \leq G_{i,t} + R_{i,t} \leq \overline{G}_{i,y} \qquad \forall i,t \tag{3.1g}$$

$$\sum_{i} R_{i,t} = 0 \qquad \forall t \tag{3.1h}$$

$$|P_{i,j,t}(\overline{P}_{k,q,y}, G_{k,t}, d_{k,t}, R_{k,t})| \leq \overline{P}_{i,j,y} \qquad \forall i,j,t \tag{3.1i}$$

$$\overline{P}_{m,n,t} = g(\overline{P}_{i,j,y}, G_{i,t}, d_{i,t}, R_{i,t}) \tag{3.1j}$$

$$\sum_{y\in b} \overline{P}_{i,j,y}l_{i,j} \leq \sum_{y\in b-1} \overline{P}_{i,j,y}l_{i,j} + z \qquad \forall b \tag{3.1k}$$

$$T_{m,n,t} = -T_{n,m,t} \qquad \forall m,n,t \tag{3.1l}$$

3.2.2 Numerical assumptions

The geographical scope of the simulation, as shown in Figure 3.1, contains a high-resolution nodal representation of the Central Western European (CWE) region, and an aggregated representation of the neighboring countries.[40] 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 actual power flows (flow-based market coupling). 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[41] 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.

The existing electricity system including power plants[42] and transmission grids[43] as of

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

[41] 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.

[42] 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).

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

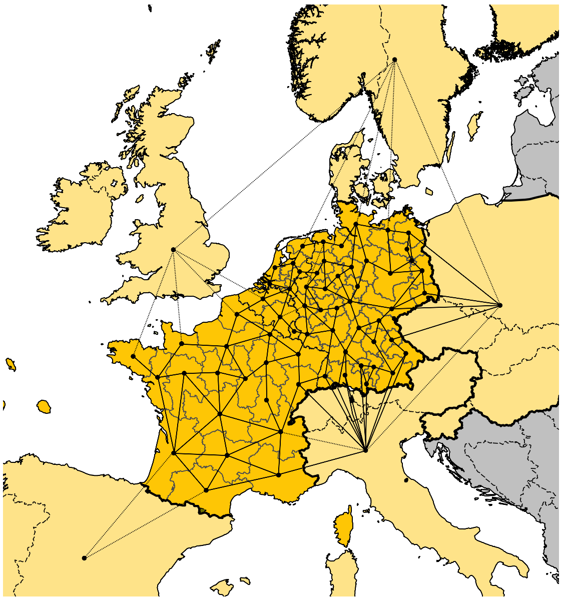


FIGURE 3.1: Representation of the CWE and neighboring regions in the model

2011 was used as the basis for the simulations of the years 2020 and 2030.[44] 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 %.[45] Years are represented by nine typical days including different demand levels, wind and solar infeed, distinguished by weekday and weekend.[46] The typical days are coupled to account for seasonal storage, and include one day to cover extreme weather events.

CO_2 emissions are constrained according to the European targets shown in Table 3.2 representing a yearly reduction of 2.2 % (compared to 2005) up to 2050 (European Commission (2014a)). If the restriction of CO_2 emissions cannot be fulfilled, the model is

[44] 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.

[45] 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 % for grid expansion projects in Germany) this seems to be a fair assumption.

[46] 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)

allowed to emit additional CO_2, for which costs of 100 EUR/t CO_2 are assumed.[47] 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 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 3.2: 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 % renewables goal in total energy consumption in Section 3. 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 levels of severity for the restriction of grid expansion, as indicated in Table 3.3. 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.

Max. grid expansion	0	5	10	15	20	30

Table 3.3: 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, load has to be curtailed to ensure technical feasibility. This would indeed be an equilibrium situation under the assumed incomplete market design. In our model, we allow to curtail

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

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.[48]

3.3 Results

3.3.1 Impacts of missing grid expansion

3.3.1.1 Redispatch and curtailed load

Figure 3.2 shows the yearly curtailed load in all scenarios, i.e., the load that could not be supplied after adjusting the dispatch with a physically feasible redispatch and grid expansion. *Scenario 0* shows the highest level of curtailed load as no grid expansion is allowed and redispatch measures are insufficient. In the CWE region, around 2 % of total load are curtailed in 2020 and nearly 3 % in 2030 if no grid expansion is allowed. All other scenarios result in load curtailment of below 0.5 % in all years. With an increase of the allowed grid expansion, less load curtailment is necessary and thus the amount of curtailed load decreases.

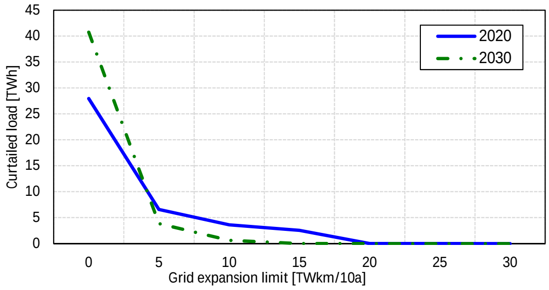


FIGURE 3.2: Curtailed load in different scenarios

Due to increasing wind capacities built up in the North of Germany without taking into the ability to transport this generation to load centers in the South, the most severe load curtailment takes place in Southern Germany. However, due to the meshed grid load also has to be curtailed in the BeNeLux-countries and the neighboring regions. Figure 3.3 shows the regional distribution and severity of load curtailment in *Scenario 0* for 2030. The distribution in the other scenarios is similar, but lower. Noticeably, load curtailment increases over time in *Scenario 0*, while it decreases in all other scenarios. This is due to the inter-temporal effect of grid expansion (cf. Section 3.3.2.2).

The overall quantity of redispatch measures shows a similar behavior over the scenarios as the curtailed load, and are highest for the most restricted case (Figure 3.4). However,

[48] Note that load curtailment is only one possible interpretation of this measure. One could also think of the corresponding results as quantities that have to be contracted outside of the market (e.g., by a reserve as in Germany).

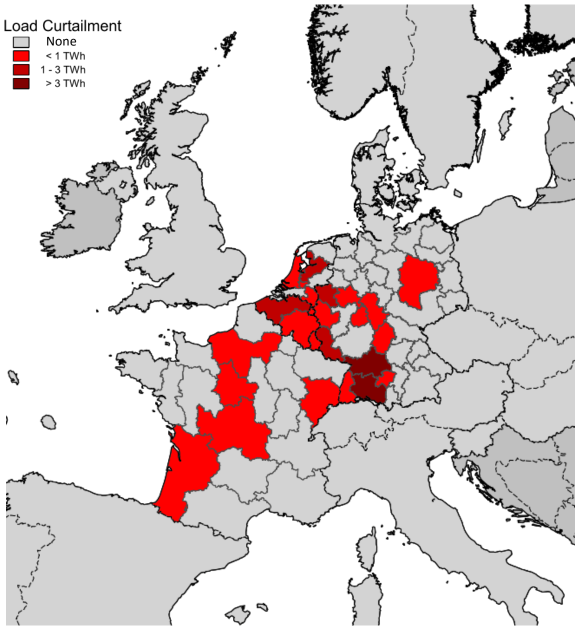


FIGURE 3.3: Geographical distribution of load curtailment in Scenario 0 for 2030

the decline in redispatch with a less restricted grid expansion is not as steep as for curtailed load. 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.

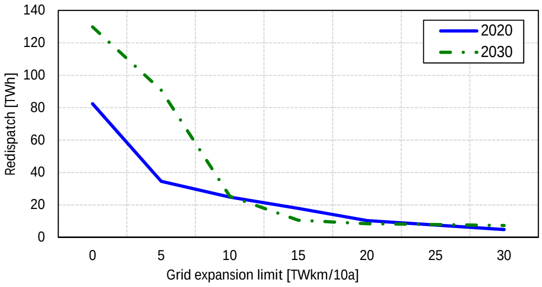


FIGURE 3.4: Redispatch measures in different scenarios

Figures 3.5 and 3.6 show the number of hours, in which transmission lines are at 100 % utilization after redispatch indicating the importance of specific transmission lines. As can be seen, the line load decreases with increasing grid expansion. However, in 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.

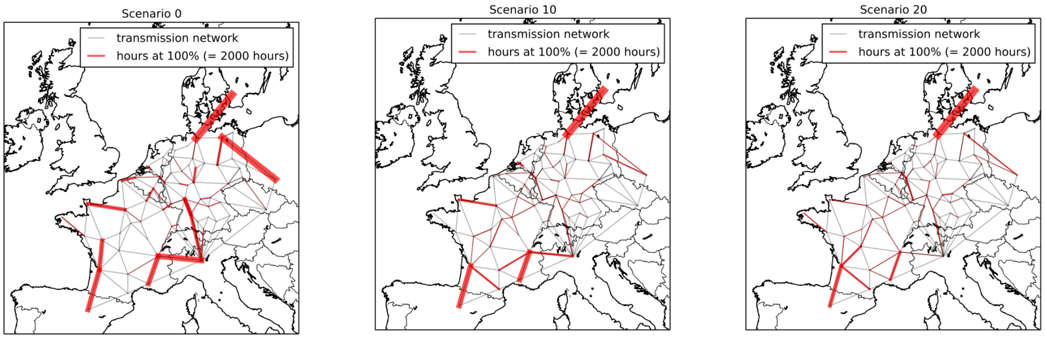


FIGURE 3.5: Line load after redispatch measures in different scenarios 2020

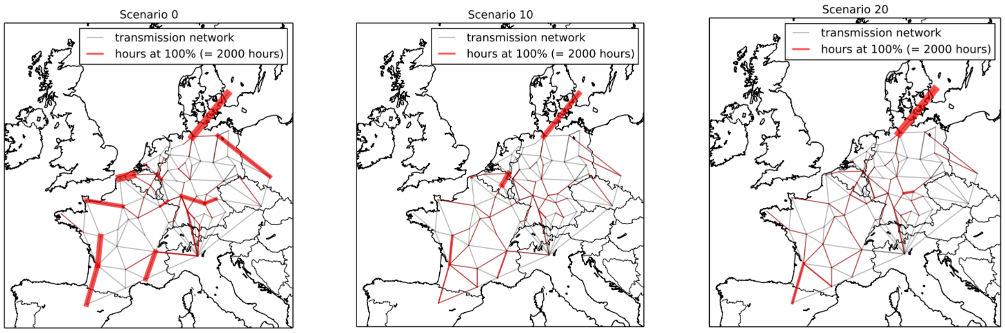


FIGURE 3.6: Line load after redispatch measures in different scenarios 2030

3.3.1.2 Fulfillment of the EU 2030 targets

In case of no grid restrictions (*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.

According to the need to curtail load, also the curtailment of renewables decreases dramatically as the restriction of grid expansion is relaxed (Figure 3.7). In 2030, 7.7 % of available PV generation is curtailed in the case of no grid expansion, which drops to just 0.4 % if 5 TWkm/10a are allowed. The drop for onshore wind from 3.3 % to 1.4 % 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.

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

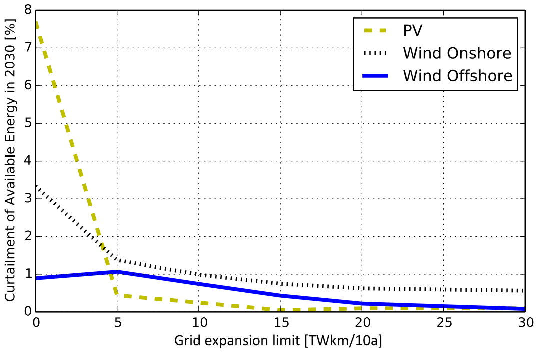


FIGURE 3.7: Curtailment of renewables for the different scenarios

is around 44 % in 2030, for *Scenario 0* the quota is about 1.5 % percentage points lower due to the curtailment.

3.3.1.3 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 % compared to the unrestricted case. Figure 3.8 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.

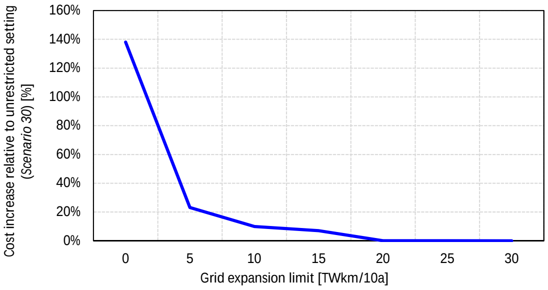


FIGURE 3.8: Total system cost decrease with grid expansion

To further analyze this result, Table 3.4 shows the composition of total systems costs (discounted to €$_{2011}$). The main variation between the scenarios results from differences in the costs of curtailed load between the scenarios where grid expansion is restricted. The increase of costs aligned to load curtailment 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, load has to be curtailed 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.

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
Load curtailment [Bn. €]	1.169.9	211.5	93.5	65.7	0	0
CO_2 shift to other sectors [Bn. €]	120	0.6	0	0	0	0
Total [Bn. €]	2238.3	1158.2	1033.9	1006.1	940.9	940.2

Table 3.4: 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.

3.3.2 Development of grid capacities

3.3.2.1 DC and AC capacities

Figure 3.9 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).

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 %.[49]

In Figure 3.9 and 3.10, 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

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

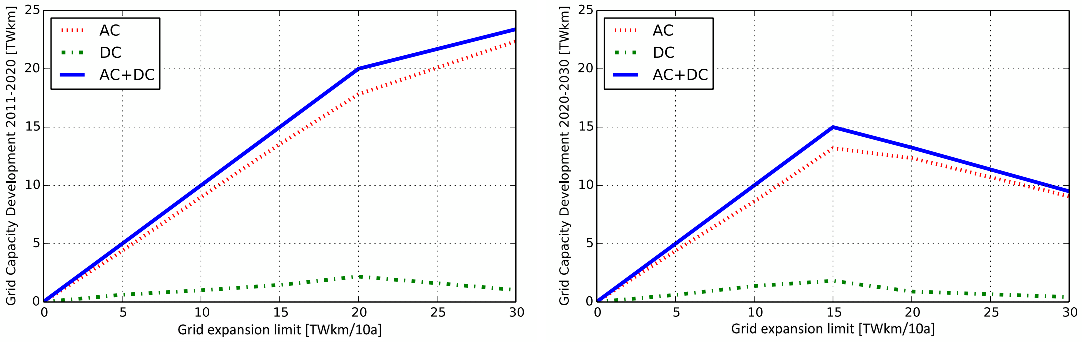


Figure 3.9: Capacity development in TWkm for the period 2011-2020 (left) and 2020-2030 (right)

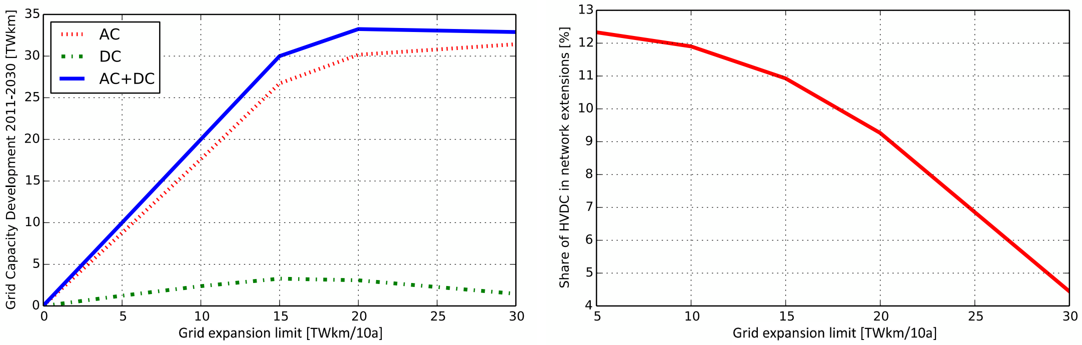


Figure 3.10: Total capacity development for 2011-2030 (left) and the share of DC in the total network expansion (right)

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. [50] However, these results also indicate 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 3.10). 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 3.9 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

[50] See Table 3.9 in the Appendix for the transmission cost assumptions

energies into the system. As a result, whenever grid restrictions are in place, DC lines allow a better system optimum.

3.3.2.2 Inter-temporal effects

In Figure 3.9, an interesting interplay between grid expansion during the two decades 2011-2020 and 2020-2030 can be seen. 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 3.10), because of the suboptimal binding grid restriction in *Scenario 20* for the decade 2011-2020.

3.3.2.3 Geographical distribution

Figure 3.11 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.

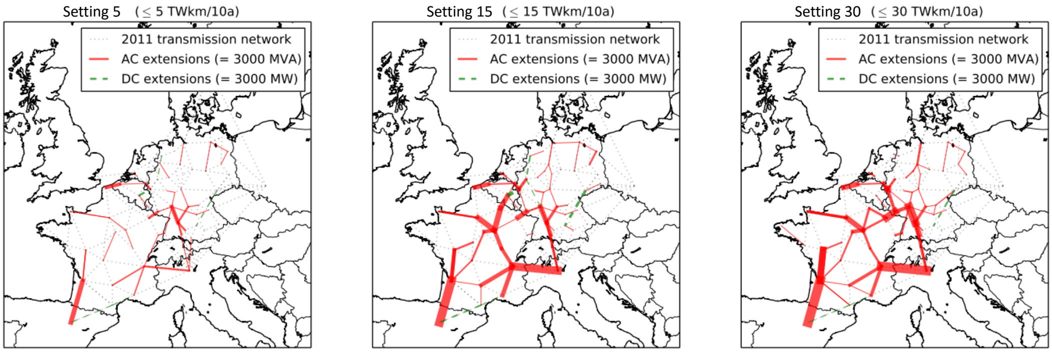

Setting 5 (≤ 5 TWkm/10a)
2011 transmission network
AC extensions (= 3000 MVA)
DC extensions (= 3000 MW)
Setting 15 (≤ 15 TWkm/10a)
2011 transmission network
AC extensions (= 3000 MVA)
DC extensions (= 3000 MW)
Setting 30 (≤ 30 TWkm/10a)
2011 transmission network
AC extensions (= 3000 MVA)
DC extensions (= 3000 MW)

FIGURE 3.11: 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*.

3.3.2.4 Inter- vs. intra-zonal grid expansion

In the grid model for the CWE region in 2011, 30 % 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 % 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 %. There is some overlap between the distribution of grid expansion calculated here and the European Commission's Projects of Common Interest[51], 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 3.11 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.

3.3.3 Generation and generation capacities

The total generation capacities and total dispatch in the CWE region in 2030 are shown in Figure 3.12 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 3.11). 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.

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

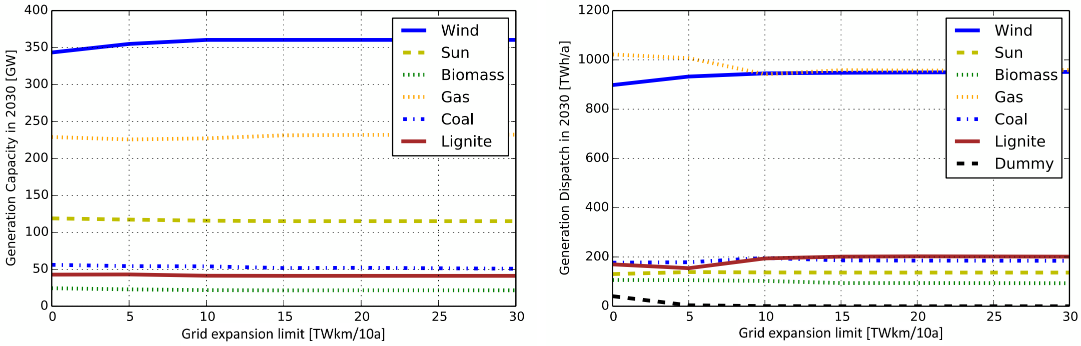


FIGURE 3.12: Total generation capacity (left) and yearly dispatch (right) in 2030 for the different grid restriction scenarios

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 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.

3.4 Conclusion

We investigated the effect of restricted grid expansion for the EU's 2030 energy strategy under the current market design. In case of grid restrictions, this market design is inherently incomplete due to the zonal markets that fail to provide efficient locational price signals. As a case study, we analyzed the development of the European electricity system with focus on the Central Western European region. The analyzed scenarios for the restricted grid expansion reach from no expansion at all to non-restricted grid expansion. To compute the outcomes, we used a linear model covering the generation as well as the transmission level with endogenous investment and dispatch decisions for both levels. Our modeling approach allowed us to explicitly represent the incompleteness of the zonal market design, which applies a flow-based market coupling along national borders with redispatch after market clearing.

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. If further measures are not taken, load has to be curtailed to ensure system stability. Although load curtailment, as a proxy for the relevance of measures in our simulation, decreases sharply if some grid expansion is allowed, we still see curtailment, even for an allowed grid expansion of 15 TWkm per decade. Affected regions are mainly those that are characterized by poor conditions for renewables and far away from (new) generation sites. Most severe load curtailment appears in Southern Germany. 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.

Generally, in the context of restricted grid expansion, an adaptation of the current market design should be considered. 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.

3.5 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 3.5: Model sets, parameters and variables

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.

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

Technology	2020	2030
Wind Onshore	1253	1188
Wind Offshore (<20m depth)	2800	2350
Wind Offshore (>20m depth)	3080	2585
Photovoltaics (roof)	1260	935
Photovoltaics (ground)	1110	785
Biomass gas	2398	2395
Biomass solid	3297	3295
Biomass gas, CHP	2597	2595
Biomass solid, CHP	3497	3493
Geothermal	10504	9500
Compressed Air Storage	1100	1100
Pump Storage	1200	1200
Lignite	1500	1500
Lignite Innovative	1600	1600
Coal	1200	1200
Coal Innovative	2025	1800
CCGT	711	711
OCGT	400	400
Nuclear	3157	3157

Table 3.7: 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 3.8: 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	1250 €/(MW*km)	6.3 €/(MW*km)
DC submarine	1100 €/(MW*km)	5.5 €/(MW*km)
DC converter pair	150000 €/MW	750.0 €/MW

Table 3.9: Assumptions for the grid extension and FOM costs

Is an inefficient transmission market better than none at all? - On zonal and nodal pricing in electricity systems

4.1 Introduction

Liberalization in electricity markets has led to an unbundling of formerly vertically integrated utilities. Consequentially, electricity generation and grid operation are tasks performed by separated entities. Still, since the balance of electricity production and consumption is crucial for system stability, generator schedules must respect the physical constraints of the transmission infrastructure to circumvent any imbalances caused by congestion. Different approaches for dealing with the transmission constraints have therefore been proposed and implemented. In some parts of the USA nodal pricing was introduced (e.g., by PJM or NYISO) providing one price for every grid node and hence explicit scarcity signals for transmission in the price differences between these nodes. Meanwhile, several countries in Europe opted for a national zone with a uniform price where transmission constraints are invisible in the market. Possible violations of the intra-zonal transmission constraints in these zones is usually being handled administratively by a Transmission System Operator (TSO). E.g., by reallocating the production of power plants (a so-called redispatch) after the market clearing with respect to congestion.

In the process of creating an internal European market for electricity, the European TSOs are obligated to deliver a review of these zones with respect to their performance. In this context, the umbrella group of the European regulators, ACER, issued a report which states that the "*European electricity target model envisages* [...] *properly defined bidding zones*", but recognizes that "[...] *the meaning of 'properly defined bidding zones' is not straightforward and needs deeper consideration.*"(ACER, 2014).

The scientific literature focusing on this topic (e.g., Harvey and Hogan (2000), Bjørndal and Jørnsten (2001)), however, essentially *is* rather straightforward, stating that the only properly defined bidding zones are actually nodes. It is argued that such a nodal pricing is superior over any zonal pricing approach, because hiding transmission scarcities from the market leads to inherent inefficiencies. A zonal aggregation could not be as efficient as a nodal pricing approach wherein the real value of transmission capacities is visible to market participants. Considering complete and competitive markets, Green (2007) provides empirical evidence for this argument for England and New Wales, and Bertsch et al. (2015) for the Central Western European region.

A more efficient market due to a larger market area with more participants is an opposed argument brought up in favor of zonal pricing. A larger market area should increase liquidity and reduce market power (e.g., CAISO (2000)). For the spot market however, Harvey and Hogan (2000) and Hogan (1999) show exemplarily that nodal pricing handles market power issues efficiently, while a zonal pricing approach might lead to perverse incentives and increasing congestion. Possible market power, but also liquidity, i.e., the possibility for traders to quickly buy or sell an asset while having a minor impact on price, is determined by the physical constraints in the spot market. A zonal pricing regime pretends more intense competition or better liquidity by allowing physically infeasible trades within a zone. Resolving these physically not viable trades brings up the need for curative measures, e.g., redispatch. In the end, market power is shifted from the spot market to the redispatch and infeasible trades, simulating liquidity, have to be administratively undone. It is not clear, why this should be (more) efficient in any case. Hogan (1999) states that an administrative pricing rule, i.e., neglecting transmission scarcities by averaging prices, does not alter the physical realities, which in the end determine the characteristics of the spot market - an argument also recognized in ACER (2014).

Empirical work (e.g., Bartholomew et al. (2003), Kristiansen (2005), Siddiqui et al. (2005), Deng et al. (2010), Adamson et al. (2010)) however, has shown that nodal pricing regimes can lack efficiency in forward markets for transmission rights. In a nodal pricing regime, energy forwards are usually traded at central hubs, while (financial) transmission rights are defined from every node to this hub. The market for these transmissions rights might not be efficient if market participants have poor expectations about the prices, transaction costs are high or liquidity is low due to few participants. If the forward market for these transmission rights is not efficient, an aggregation of nodes to zones and hence, a reduced number of transmission rights to be traded, might be advantageous. An aggregation of nodes to bidding zones implicitly hedges all risks of transmission constraints and socializes the costs via the curative measures in the spot market. The remaining question and the issue addressed in this paper is therefore: Under which circumstances does the inefficiency of a transmission forward market (in a nodal pricing regime) matter more than the inefficiencies induced by neglecting transmission (in a zonal pricing regime)?

To show the general effects of the two inefficiencies of the different pricing regimes, a simple two node model with two producers, a retailer and a transmission system operator with a spot and forward market is developed. On the forward market a transmission right for hedging against the risk of congestion can be traded, while on the spot market only energy is considered. The TSO clears the market in a welfare optimal way while ensuring physical feasibility. For performing comparative statics of nodal and zonal pricing, a more complex model incorporating more nodes, loop flows as well as energy and transmission forwards is used. The spot and forward market model by Bessembinder and Lemmon (2002) and de Maere d'Aertrycke and Smeers (2013) serves as a base. The stochastic equilibrium model by de Maere d'Aertrycke and Smeers (2013) incorporates missing liquidity in a nodal pricing system via a volume constraint for forward transmission contracts. I extend their model by a zonal pricing approach and a producer-only redispatch. Furthermore, an exogenous bid-ask-spread to model effects of inefficiencies such as missing liquidity, transaction costs etc. is implemented. Compared to a volume-constrained approach, this allows to solve for unique prices of transmission forward contracts and hence, a unique equilibrium, which is essential for comparing the two pricing regimes. A consistent numerical setting (proposed by Chao and Peck (1998)) is chosen and a systematic numerical analysis performed by varying the influencing fundamental factors such as grid restrictions, supply and demand properties as well as the bid-ask spread and risk aversion.

The contribution to the literature is twofold: First, a consistent model for comparing the effects of inefficiencies in nodal and zonal pricing regimes is developed. Second, comparative statics for an established numerical framework is performed to identify relevant circumstances regarding the comparison of both pricing regimes.

With no inefficiencies in transmission forward markets, the arguments of the theoretical literature can be confirmed and it can be concluded that nodal pricing performs always better than zonal pricing. The relative performance of zonal pricing improves with decreasing congestion and is equal to nodal pricing in case of no congestion at all. In addition, smaller supply cost differences and a more inelastic demand work in favor of zonal pricing. In the case of an inefficient transmission forward market, situations appear wherein zonal pricing performs better than nodal pricing. The total number of such cases increases with increasing bid-ask-spread and risk aversion.

The results show that efficiency of transmission right markets does in fact matter. Hence, for finding 'properly defined bidding zones', the efficiency of forward markets for transmission should be one criterion. With a zonal pricing in place this could mean that, e.g., the geographical coverage of a zone should be defined as a trade-off between the inefficiencies of a zonal pricing and a nodal pricing regime. The respective measure or a possible change of pricing mechanisms, however, has to include political considerations and transaction costs of system adaption.

The paper is organized as follows: Section 2 introduces a simple model. The model used for comparative statics, the numerical framework and results are presented in Section 3. Section 4 concludes.

4.2 A simple model

Within this section, a simple two node model for comparing a zonal pricing regime with a redispatch of producers and a nodal pricing regime with an inefficient transmission forward market is established. This allows to identify some general effects, which also apply to larger models. At first, the differences in the spot market outcomes with and without congestion are shown. Next, the differences considering a an additional forward market are analyzed.

4.2.1 Spot market outcomes without congestion

Consider an electricity system with two nodes and a single time period. A producer with constant marginal costs of c_1 is located at node 1 and a producer with constant marginal costs of c_2 is located at node 2. Both producers have infinite production capacities, but producer 2 has higher costs, i.e., $c_1 < c_2$. The final consumer demand is located at node 2 and represented by some demand function $D^l(p^r)$ with $\frac{\partial D^l(p^r)}{\partial p^r} < 0$. A retailer buys electricity at price p^s from the producers and sells it to the final consumers at price p^r, which is the (weighted) average price of all possible spot market realizations $p^r = E(p^s)$.[52] Producers and retailers are considered as price-takers. The transmission capacity between the nodes k is considered to be larger than the maximum demand, i.e., no congestion occurs. With no congestion, the market outcomes for nodal (NP) and zonal pricing (ZP) are the same as indicated in Figure 4.1. Only the producer at node 1 is dispatched, the spot price is $p^s = c_1$ and the retail price is $p^r = p^s = c_1$ resulting in a demand of $q^* = D^l(c_1)$.

Profits are zero for producers and the retailer. Hence, total welfare for the case with no congestion (noc) is determined only by the consumer rent:

$$W_{noc}^{NP,s} = W_{noc}^{ZP,s} = \int_{c_1}^{\infty} D^l(p^r)dp^r \tag{4.1}$$

4.2.2 Spot market outcomes with congestion

Now consider that transmission capacity is limited due to a higher demand (i.e., $k < D^h(c_2)$). For ensuring physical feasibility, a benevolent transmission system operator (TSO) is introduced. In the zonal pricing regime, the TSO performs a redispatch of the producers after the market outcome until the transmission constraints are fulfilled. In the nodal pricing regime, the TSO clears the market considering the transmission constraint. The spot market outcomes now differ for the pricing regimes.

[52] In this case there is only one realization and therefore $p^r = p^s$, but generally there is more than one possible spot market outcome as will be shown later.

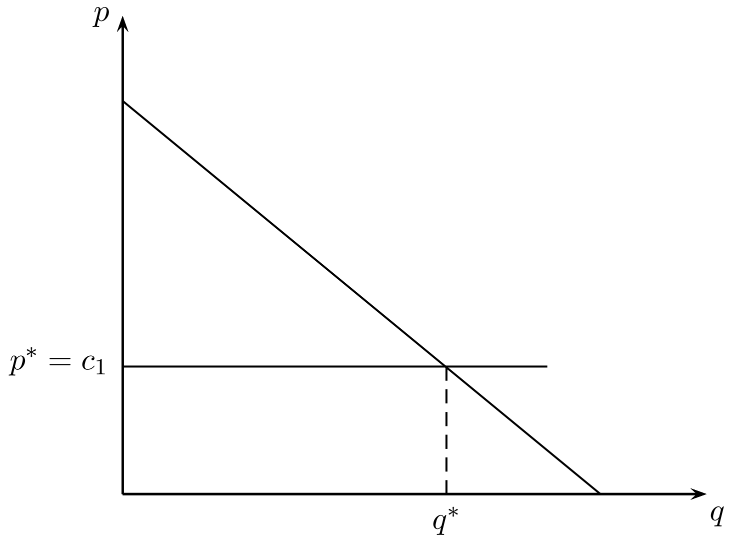


Figure 4.1: Spot market result without congestion

4.2.2.1 Nodal pricing outcome

The TSO clears the market with a cost-minimal dispatch considering the transmission constraint. The producer at node 1 now generates electricity up to k, whereas the producer at node 2 produces from k to $q^* = D^h(c_2)$ (Figure 4.2). The prices reflect the marginal costs at the respective nodes with $p_1^s = c_1$ and $p^r = p_2^s = c_2$.

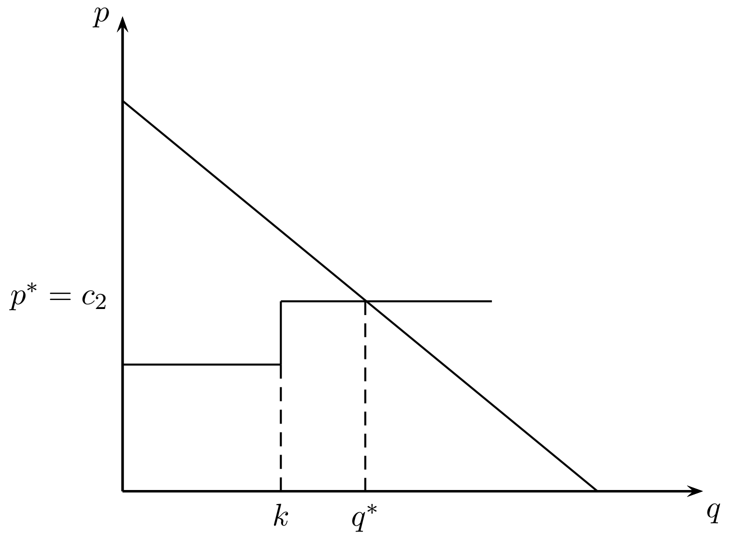


Figure 4.2: Spot market result for nodal pricing with congestion

It is assumed that the TSO receives the profit resulting from the price difference between the nodes and the quantity transmitted from node 1 to node 2. Profits of the TSO and welfare in this congested case (con) are

$$\pi_{tso}^{NP,s} = (c_2 - c_1)k \tag{4.2}$$

$$W_{con}^{NP,s} = \pi_{tso}^{NP,s} + \int_{c_2}^{\infty} D^h(p^r)dp^r \tag{4.3}$$

4.2.2.2 Zonal pricing outcome

In the zonal pricing regime, the market outcome is calculated as in the case of no congestion, i.e., prices are $p^r = p^s = c_1$, demand is $q^* = D^h(c_1)$ and profits are zero for producers and the retailer. However, this outcome is technically not feasible, because the producer at node 1 would export more than k to node 2. The TSO now redispatches the two producers until technical feasibility is achieved. For this, the TSO reduces the quantity of the producer at node 1 to k and increases the production of the producer at node 2 to $D(c_1) - k$. In Figure 4.3 the left graph shows the spot market outcome and the right graph shows the real dispatch. For the quantity from k to q^* the producer at node 2 is dispatched, but still, electricity is traded at $p^s = c_1$.

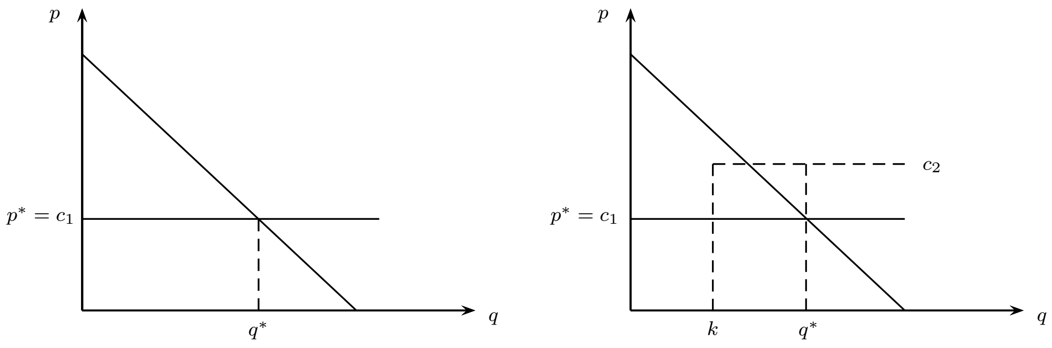


FIGURE 4.3: Spot market result for zonal pricing with congestion

The additional costs result in a negative TSO profit. Profit and welfare are

$$\pi_{tso}^{ZP,s} = (c_1 - c_2)(D^h(c_1) - k) \tag{4.4}$$

$$W_{con}^{ZP,s} = \pi_{tso}^{ZP,s} + \int_{c_1}^{\infty} D^h(p^r)dp^r \tag{4.5}$$

4.2.2.3 Comparing the spot market outcomes

Comparing $W_{con}^{NP,s}$ and $W_{con}^{ZP,s}$ yields:

$$\begin{aligned}\Delta W_{con} &= W_{con}^{NP,s} - W_{con}^{ZP,s} \\ &= (c_2 - c_1)k + \int_{c_2}^{\infty} D^h(p^r)dp^r - (c_1 - c_2)(D^h(c_1) - k) - \int_{c_1}^{\infty} D^h(p^r)dp^r \\ &= \underbrace{\int_{c_2}^{c_1} D^h(p^r)dp^r}_{<0} + \underbrace{(c_2 - c_1)D^h(c_1)}_{>0} \qquad (4.6)\end{aligned}$$

and see immediately that ΔW has to be always larger than zero if $\frac{\partial D^l(p^r)}{\partial p^r} < 0$, i.e., if demand decreases in price. In the zonal pricing market outcome demand is too high due to neglecting the transmission capacities. This inherent inefficiency is induced by only allowing producers for redispatch. In Figure 4.4 the striped area marks the welfare delta induced by the inefficiency.

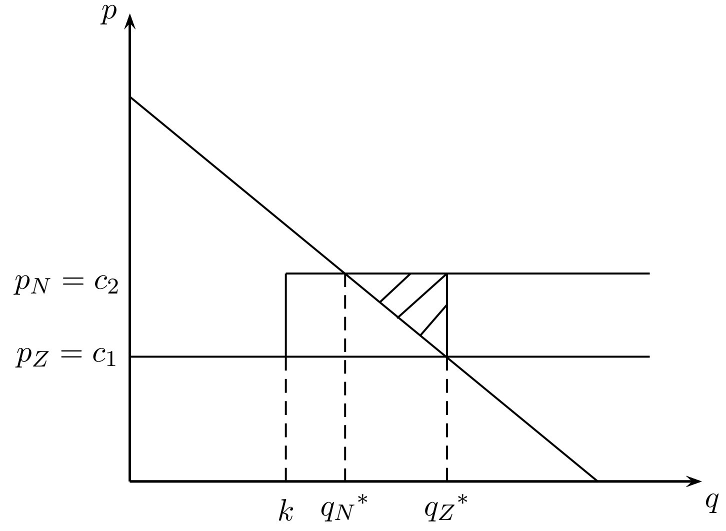


FIGURE 4.4: Comparison of spot market results with congestion

If demand is allowed for redispatch, both regimes would lead to the same outcome. The same would be the case, if demand is inelastic. This means the more inelastic demand and the more negligible the cost differences of the nodes, the smaller is the welfare delta between the two regimes.

4.2.3 Forward market outcome

A forward market is considered, on which a transmission forward can be traded. The transmission forward or financial transmission right (FTR) is defined as the right to collect the congestion rent, i.e., the price differences, between two nodes or zones. There are two possible spot market outcomes at the time trade takes place at the forward market, namely the ones described above with either no congestion (*noc*) (probability μ) or congestion *con* (probability $1-\mu$). Empirical literature (e.g., Viehmann (2011)) has found that market participants in electricity markets show risk averse behavior. Hence, producers and and the retailer are assumed to be risk averse in the sense that they require additional capital for possible losses. The TSO is considered to be risk neutral.[53]

In zonal pricing, the market participants do not see any transmission capacity in the spot market outcome and hence, there is no forward market for transmission. The prices p^s and p^r are always the same, regardless of congestion. Producers and retailers -despite being risk averse- have no desire to hedge any possible spot market realization due to the variance of their profits being zero. Overall welfare in the forward market for zonal pricing is:

$$W^{ZP,f} = \mu W_{noc}^{ZP,s} + (1-\mu) W_{con}^{ZP,s} = \mu \left[\int_{c_1}^{\infty} D^l(p^r) dp^r \right] + (1-\mu) \left[\pi_{tso}^s + \int_{c_1}^{\infty} D^h(p^r) dp^r \right] \tag{4.7}$$

In the nodal pricing regime, the forward market outcome is different. The producer at node 1 gets always the same price regardless of congestion. The producer at node 2 and the retailer however, face two possible price realizations. The expectation value of the profits are still zero, but the retailer has a loss with probability $1-\mu$ due to the fixed retail price. Risk aversion of the retailer is indicated by some function $U(\pi)$ describing the additional capital requirements of losses with $U(\pi) < \pi \ \forall \ \pi < 0$ and $U(\pi) = \pi \ \forall \ \pi \geq 0$. Hence, the retailer is willing to give up some profit from the case of no congestion to hedge against the possible losses. This can be done with the TSO offering a FTR x at price $p^{f,bid}$ which allows the consumption of the congestion rent. The retailer can buy this FTR at price $p^{f,ask}$. The profit functions change to

$$\pi_r^{NP,f} = \mu U(\pi_{r,noc}^s) + (1-\mu) U(\pi_{r,con}^s) + p^{f,ask} x \tag{4.8}$$

$$\pi_{tso}^{NP,f} = (1-\mu) \pi_{tso,con} - p^{f,bid} x \tag{4.9}$$

The FTR is independent of the realization of either the congested or the uncongested case since it is an option to consume the congestion rent in case of congestion. The risk averse retailer is willing to give up some profit from the uncongested case in order to reduce her losses in the congested case. The risk neutral TSO will take over all risk of the retailer, if $p^{f,ask} = p^{f,bid}$. The bid-ask-spread is equal to zero if there are no transaction costs and the traded contract is fully liquid. However, it can be greater than

[53] Of course, also a risk averse TSO could be assumed. The reason for the assumption of risk neutrality is due to making the pricing regimes comparable: In the zonal pricing regime, neglecting transmission capacities could be interpreted as a risk neutral administrative hedge against congestion risks. Hence, an equivalent assumption becomes necessary in the nodal pricing regime.

zero if the asset lacks liquidity or transaction costs occur, i.e., $p^{f,ask} < p^{f,bid}$. In this case, the TSO is not able to fully take over the risk of the retailer, who is left with some unhedged losses.

Overall welfare in the forward market for nodal pricing is:

$$\begin{aligned} W^{NP,f} =& \mu \left[U(\pi^{s}_{r,noc}) + \int_{c_1}^{\infty} D^{l}(p^{r}) dp^{r} \right] \\ &+ (1-\mu) \left[U(\pi^{s}_{r,con}) + \pi^{s}_{tso} + \int_{c_2}^{\infty} D^{h}(p^{r}) dp^{r} \right] - p^{f,bid}x + p^{f,ask}x \end{aligned} \tag{4.10}$$

The comparison of the welfare in the forward markets yields:

$$\Delta W^{f} = W^{NP,f} - W^{ZP,f} = \Delta W_{con} + \underbrace{(1-\mu) \left[U(\pi^{NP,s}_{r,con}) + \pi^{NP,s}_{tso,con} \right]}_{\text{Costs of unhedged risks}} - \underbrace{(p^{f,bid} - p^{f,ask})x}_{\text{Bid-ask-spread inefficiency}} \tag{4.11}$$

In addition to the spot market welfare delta, there are two new parts stemming from the possibly unhedged loss of the retailer and the possible bid-ask-spread. Both terms work in favor of zonal pricing since they can at largest be zero, if $p^{f,ask} = p^{f,bid}$. In case the bid-ask-spread is larger than zero, the retailer cannot fully hedge the losses by buying FTRs from the TSO. Hence, both terms become negative and reduce the delta between nodal and zonal pricing. In case of a positive bid-ask-spread the question which regime performs better depends on whether this inefficiency is larger than the inefficiency induced by the producer-only redispatch of the zonal pricing regime, included in ΔW_{con}.

From the analysis of the spot market outcome it was already obtained that smaller cost differences and a more inelastic demand improve the relative performance of zonal compared to nodal pricing. From the forward market analysis it can be concluded that a higher bid-ask-spread and higher costs of unhedged risks work in the same direction and may even (over)compensate the inefficiencies of zonal pricing in the spot market.

4.3 Comparative statics

In order to further investigate the trade-off obtained in the simple model, more details of the electricity grid are now considered. There are two main characteristics which deserve special consideration, namely the spatial distribution and correlation of demand as well as different grid configurations. Flows in electricity grid can usually not be directed from node to another and hence, an injection of electricity anywhere in the grid impacts electricity flows on all lines. This becomes especially relevant in combination with changing demand patterns. Therefore, a model incorporating more nodes, the spatial distribution of demand, loop flows, more possible spot market outcomes and a forward market for transmission rights *and* energy is developed. With this model, a numerical analysis with a broad variation for the named characteristics is performed. The often applied setting by Chao and Peck (1998) is used as consistent numerical base.

4.3.1 A slightly more complicated model

For the general formulation of the model I follow Bessembinder and Lemmon (2002) and de Maere d'Aertrycke and Smeers (2013) and extend their nodal pricing formulation to incorporate zonal pricing with a producer-only redispatch. Furthermore, an alternative modeling approach for ineffícency of the transmission forward market is chosen by implementing a bid-ask-spread instead of constraining the trading volume of transmission forwards.

4.3.1.1 Market participants and TSO

The market participants are denoted by N consisting of producers N_P and retailers N_R. There is either one producer or one retailer at any node ν. Furthermore, an arbitrary number of nodes can be in one market area m, n of all markets M. Producers have a cost function consisting of a fixed component a and variable costs b depending on the quantity sold in the spot market area. In line with Bessembinder and Lemmon (2002) a quadratic cost function is assumed.

$$c_\nu(q_\nu) = a_\nu q_\nu + \frac{b_\nu}{2} q_\nu^2 \tag{4.12}$$

Producers sell their production in the spot market at price p_m^s and hence their profit is

$$\pi_\nu^s = p_m^s q_\nu - c_\nu(q_\nu). \tag{4.13}$$

Retailers buy production at the spot market price p_m^s and sell it to the consumers at a fixed price p_m^r. Their profit is

$$\pi_\nu^s = (p_m^r - p_m^s) q_\nu. \tag{4.14}$$

The linear inverse demand function $p_\nu(q_\nu)$ of the consumers at node ν is subject to an exogenous shock caused by, e.g., changing weather conditions.[54] The shocks are indicated in the demand function by realizations ω of the exogenous random variable a_ν.

$$p_\nu(q_\nu) = a_\nu^\omega - b_\nu q_\nu. \tag{4.15}$$

[54] This basically corresponds to the different cases in the simple model.

The fixed price p_m^r is assumed to be the average price resulting from the market clearing with all possible realizations of a_ν plus some markup, i.e., $p_m^r = \sum_\omega^\Omega prob^\omega p_m^\omega + \mu$ with $\sum_\omega prob^\omega = 1$. The markup is the profit the retailer gets from selling electricity to the final consumers. To a certain extent the markup works as an insurance for having a fixed retail price and a volatile spot price. Retailer profit can be negative for some realizations a_ν^ω depending on the properties of the random variable a_ν and the markup. To hedge themselves against the volatility of the spot market and possible losses, producers and retailers have the possibility to buy or sell contracts $c \in C$ with a price p_c^f and a quantity $x_{c,\nu}$ on the forward market.[55] The underlyings of the forward contracts can either be based on energy sold in the spot market (p_m^s, q_m^s) or on the right to collect congestion rent, which is defined as the price difference and the traded quantities between two spot markets $(\Delta p_{m,n}^s, \Delta q_{m,n}^s)$.[56] The right to collect congestion rent corresponds to a financial transmission right. The forward prices are $p_c^{f,bid}$ for going long and $p_c^{f,ask}$ for going short. The corresponding forward contract quantities $x_{c,\nu}^{bid}$ and $x_{c,\nu}^{ask}$ are not dependent on the realization ω. The difference between the prices indicates the bid-ask-spread. This differs from the formulation in de Maere d'Aertrycke and Smeers (2013), where the tradeable volume was limited to represent an inefficient transmission forward market. Treating this inefficiency as a bid-ask-spread yields the nice property that the forward price of each contract for bid or ask is equal for all players. For the later solution this implies that only one equilibrium exists, which is then arbitrage free if the bid-ask-spread is exogenous.[57]

The profit of a producer or retailer in the forward market for one possible realization of the spot market is

$$\Pi_\nu^\omega = \sum_{c=1}^{C}((p_c^{s,\omega} - p_c^{f,bid})x_{c,\nu}^{bid} + (p_c^{f,ask} - p_c^{s,\omega})x_{c,\nu}^{ask}) + \pi_\nu^{s,\omega} \tag{4.16}$$

The profit realization Π_ν^ω in the forward market depends on the loss or profit in the spot market $\pi_\nu^{s,\omega}$. Possible losses in the spot market is costly in the sense of some sort of additional capital requirement. Hence, market participants try to hedge against losses in the spot market by giving up some possible profits.[58] A producer or retailer optimizes the overall profit in the forward market by buying forward contracts considering the

[55] Note that the prices are the same for all market participants while the quantities are individual.

[56] Due to loop flows the traded quantities are not necessarily the physical flows!

[57] This is not shown analytically, but the numerical analysis and the behavior of the solution algorithm for a wide range of starting points indicate that the equilibrium is unique.

[58] For modeling this, de Maere d'Aertrycke and Smeers (2013) propose a weighted sum of the expectation of losses and a conditional value at risk (CVaR) as a coherent risk measure (E-CVaR). The conditional value at risk defines the expected value above some value at risk. I follow their definition, but drop the time variable since only one time period is considered: $\text{E-CVaR}_{\alpha,\beta} = (1-\beta)\mathbb{E}[-\Pi_\nu] + \beta\text{CVaR}_\alpha(\Pi_\nu)$.

additional costs of possible losses indicated by $U_\nu^\omega(\Pi_\nu^\omega)$:[59]

$$\max_{x_\nu} \left\{ \sum_\omega \text{prob}^\omega \left[(1-\beta_\nu)\Pi_\nu^\omega - \beta_\nu \alpha_\nu^{-1} U_\nu^\omega(\Pi_\nu^\omega) \right] \right\} \tag{4.17}$$

The weight β defines the relative importance of expected losses versus the additional capital requirements needed for the average losses above some value at risk (specified by the quantile above α). This means a lower α requires more risk capital and a higher β puts more emphasis on these capital requirements.

As stated in the simple model, the TSO has some profit in the spot market due to the collected congestion rent, i.e., the price differences times the traded energy minus the redispatch costs R.

$$\pi_{tso}^s = \frac{1}{2} \sum_m^M \sum_n^M |\Delta p_{m,n}^s||q_{m,n}| - R. \tag{4.18}$$

Considering the forward market, the TSO has the additional role to emit financial transmission rights and to act as the ultimate counter-party for trading these contracts. Emission is limited by the actual transmission capacities available in the spot market. The TSO is only allowed to trade FTRs but no energy forward contracts, leading to the profit for each realization of the spot market:

$$\Pi_{tso}^\omega = \sum_{c=1, c\notin c_e}^C \left((p_c^{s,\omega} - p_c^{f,bid}) x_{c,tso}^{bid} + (p_c^{f,ask} - p_c^{s,\omega}) x_{c,tso}^{ask} \right) + \pi_{tso}^{s,\omega} \tag{4.19}$$

The optimization problem for the TSO is the same as for the other market participants, stated in Equation 4.17.

4.3.1.2 The spot market

The forward and spot market are sequential. Producers, Retailers and the TSO do not know the exact market outcome of the spot market, but do know all possible outcomes and the respective probability distribution. Furthermore, a competitive spot and forward market is considered, with producers and retailers as price takers and the TSO as the welfare-maximizing market clearer. This allows to first compute all possible spot market outcomes and then solve for the equilibrium on the forward market.

[59] Function $U_\nu^\omega(\Pi_\nu^\omega)$ is defined as in the simple model, i.e., $U_\nu^\omega(\Pi_\nu^\omega) < \Pi_\nu^\omega \ \forall \ \Pi_\nu^\omega < 0$ and $U_\nu^\omega(\Pi_\nu^\omega) = \Pi_\nu^\omega \ \forall \ \Pi_\nu^\omega \geq 0$.

Spot market clearing

$$\max_{q_\nu} \left[\sum_{\nu \in N^R} \int_0^{q_\nu} p_\nu(q_\nu) dq_\nu - \sum_{\nu \in N^P} \int_0^{q_\nu} c_\nu(q_\nu) dq_\nu \right] \tag{4.20a}$$

$$\text{s.t.} \sum_{\nu \in m} q_\nu + \sum_{-m} (q_{m,-m} - q_{-m,m}) = 0 \qquad \forall m \in M \tag{4.20b}$$

$$f(q_{m,-m}) \leq k_{m,-m} \qquad \forall m \in M \tag{4.20c}$$

$$q_\nu \geq 0 \tag{4.20d}$$

The spot market clearing is done by the TSO by maximizing the welfare over quantities q (Equation 4.20a) subject to the market balance (Equation 4.20b) and the transmission constraints (Equation 4.20c). If ν is identical to m (implying that real transmission scarcities are visible in the market), the spot market outcome is at the same time the optimal dispatch and the problem is physically feasible. This setting corresponds to a nodal pricing regime. The function f then maps power injections and withdrawals to load flows on lines. If ν is not identical to m and hence, transmission scarcities are only incompletely considered, f represents trade flows, restricted by some trading capacity offered to the market. This corresponds to a zonal pricing regime, wherein the zone definition depends on the mapping of nodes to markets. A subsequent optimization to ensure physical feasibility has to be performed.[60] The problem is formulated as cost minimization representing a redispatch of power plants which does *not* alter the profits obtained in the spot market. One implicit assumption is that the TSO has full information about the cost functions of the producers.

Redispatch

$$\min_{q_{\nu \in N^P}} R = \left[\sum_{\nu \in N^P} \int_0^{q_\nu} c_\nu(q_\nu) dq_\nu \right] \tag{4.21a}$$

$$\text{s.t.} \qquad \sum_{\nu \in N^P} q_\nu - \sum_{\nu \in N^R} q_\nu = 0 \tag{4.21b}$$

$$f(q_\nu) \leq k_{\nu,\nu'} \tag{4.21c}$$

The system balance constraint (equation 4.21b) is defined over all zones, i.e., implying a coordinated cross-border redispatch. Equation 4.21c now represents the real transmission constraints by indexing over ν instead of m. The objective value of this cost minimization is given by the redispatch costs R, which are part of the TSO's profit function (Equation 4.18).

[60] The sequential optimization approach ensures that the TSO has no means to actively profit from redispatching power plants.

The total welfare of the spot market is computed as the sum of producer, retailer and TSO profits plus the end consumer rent retrieved over the demand function.

$$W^s = \sum_{\nu \in N} \pi^s_\nu + \pi^s_{so} + cr^r \tag{4.22}$$

4.3.1.3 The forward market

The forward market is cleared before the realization of the spot market. On the forward market the market participants optimize their profits as stated before. The TSO emits FTRs and acts as the ultimate counterparty for transmission forward contracts. The market is cleared if no further forward quantity is traded, i.e., retailers, producers and the TSO have optimized their forward contracting. The forward market clearing can be defined by the sum of the single optimization problems plus some market clearing and bid-ask-spread constraint. Index ν is slightly abused by including the TSO to reduce the complexity of the formulation.

Forward market clearing

$$\max_{x_\nu} \sum_{\nu \in \{N,TSO\}} \left\{ \sum_{\omega} \text{prob}^\omega \left[(1-\beta_\nu)\Pi^\omega_\nu - \beta_\nu \alpha_\nu^{-1} U^\omega_\nu \right] \right\} \tag{4.23a}$$

$$\text{s.t.} \quad \sum_\nu x^{bid}_{\nu,c} - \sum_\nu x^{ask}_{\nu,c} = 0 \tag{4.23b}$$

$$f(x^{bid}_{tso,c} - x^{ask}_{tso,c}) \leq k_{m,n} \tag{4.23c}$$

$$p^{bid}_c - p^{ask}_c \geq \chi_c \tag{4.23d}$$

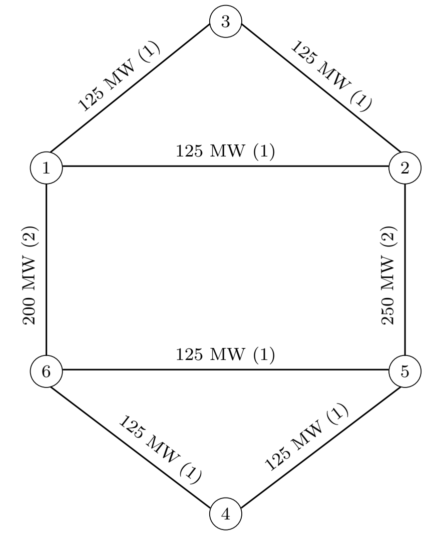
While Equation 4.23a simply sums up the optimization problems of the single players, Equation 4.23b represents the market clearing condition for all forward contracts. Equation 4.23c guarantees feasibility of the transmission constraints in the spot market, i.e., restricts the TSO's emission of FTRs to feasible quantities. As in the spot market, the function f describes a mapping of traded quantities to physical flows, which depends on the nodal or zonal pricing regime. Considering nodal and zonal pricing the main difference between the regimes lays in the different number of forward contracts needed. If in the nodal pricing regime some hubs corresponding to the zones in the zonal pricing are defined, the number of energy contracts might be the same. The number of transmission contracts however, is different. In a nodal pricing regime, there are as many transmission contracts as nodes minus the hub. With zonal pricing, the number of transmission contracts corresponds to the number of connected zones. Equation 4.23d introduces a bid-ask-spread, exogenously defined by some value χ.

Total welfare of the forward market can be computed considering the profit, the additional capital requirement from unhedged risk and the consumer rent:

$$W^f = \sum_{\omega} \text{prob}_{\omega} \left[\sum_{\nu \in \{N,TSO\}} (\Pi_{\nu,\omega} - \alpha_{\nu}^{-1} U_{\nu,\omega}) + cr^{r,\omega} \right] \tag{4.24}$$

4.3.2 Numerical analysis

For the numerical analysis the six-node spot market model introduced by Chao and Peck (1998) is applied. For the forward market the general setting from de Maere d'Aertrycke and Smeers (2013), who also use the Chao-Peck model as a basis, is taken. The spot market model has 3 production and 3 retailer nodes. The structure and the standard supply and demand functions for these nodes can be seen in Figure 4.5.



Node	Supply functions $C_v(q_v)$
1	$10 + 0.05q$
2	$15 + 0.05q$
4	$42.5 + 0.025q$

Node	Demand functions $p_v(q_v)$
3	$37.5 - 0.05q$
5	$75 - 0.1q$
6	$80 - 0.1q$

FIGURE 4.5: Basic spot market model (Chao and Peck, 1998), taken from de Maere d'Aertrycke and Smeers (2013)

In the nodal pricing regime, each node represents a spot market, while trading between these markets is restricted by transmission constraints. For the zonal pricing regime all nodes are aggregated into one spot market. This allows us to ignore transmission capacity allocations, which is a rather difficult issue. Due to loop flows the available transmission capacity depends on the actual dispatch. Hence, the transmission capacity offered to the market has to be valid for all possible dispatch situations. Furthermore, contingencies are included when assessing the possible transmission capacity, making it even more difficult to come up with a reasonable value. One should keep in mind that this configuration is the worst possible case for zonal pricing and could be relaxed by, e.g., defining two zones. For our purpose this means that if zonal pricing performs better in any case, this could also happen with better configuration of bidding zones. To ensure physical feasibility, a cost-based redispatch of power plants -as described above- is applied, which does not affect the profit of the market participants. A risk-neutral TSO is assumed, i.e, $\beta = 0$. This means that in the zonal pricing regime, all congestion costs are present in the actual redispatch (and do not appear in the forward market).

Drivers for the behavior of market participants in the forward market are costs of risk, prices and profits. The latter two depend fundamentally on supply and demand characteristics, grid configuration and markup. Furthermore, the bid-ask-spread impacts the achievable welfare of any market. All the mentioned parameters are varied as shown in Table 4.1.

Table 4.1: Parameter variations

	Parameter	**Variation**	**Steps**	**Scenarios**
Demand	a_{ν}^{ω}	fully correlated, semi-correlated and non-correlated (with different variance)		4
Supply	a_{ν}	between 20% and 180% at each node	40%	5
Grid	k	Line 1-6 between 0 and 400 Line 2-5 between 50 and 450	50	9
Markup	μ	between 0 and 20% for each retailer	0.05	5
Risk aversion	α	0.1, 0.2, 0.4 and 0.8		4
	β	0.1, 0.5 and 0.9		3
Bid-ask-spread	χ	0, 0.1, 0.25, 0.5, 1, 2, 4, 6		8

For each scenario, different dispatch realizations (indicated by ω, $\sum_{\omega} prob^{\omega} = 1$) with a varying demand a_{ν}^{ω} are calculated. In the dispatch realizations, the axis intercept of the demand function is varied between 75% and 125% (respectively 70% and 130% for the high variance - no correlation scenario) in steps of 12.5 % (30 % for the high variance - no correlation scenario) at each demand node. The correlation of demand levels at nodes is differentiated in four demand scenarios, leading to 202 different dispatch situations in total.[61] Beside the demand correlation, additional scenarios consist of variation of supply costs and grid capacities. There are 180 spot market scenarios with a total of 9,090 different dispatch realizations. For the forward market, scenarios with variation of the markup of retailers, the costs of risk (α,β) and the bid-ask-spread are calculated. Together with the spot market scenarios this adds up to 86,400 scenarios in total.

[61] For each demand scenario, a different number of dispatch realizations is needed, but their probability always adds up to 1. In the no correlation case, all 125 possible combinations are considered, in the no correlation - high variance case all 27 combinations are considered. The semi-correlated scenario adds up to 45 and the fully correlated scenario to 5 dispatch situations.

The dispatch realization can be computed via an integrated maximization approach for the nodal pricing regime, while for the zonal pricing regime, the additional redispatch has to be calculated. For solving the forward market, I follow the solution algorithms suggested by de Maere d'Aertrycke and Smeers (2013). They solve the forward market as an iterative linear problem by fixing the forward price and updating it with the marginal of the equilibrium constraint. The iterative solution algorithm allows us to introduce the bid-ask-spread χ as the difference between the bid and the ask price in the updating of the forward prices. With (bid and ask) prices being equal for all market participants the resulting equilibrium is a global optimum.[62] The global optimum property enables the comparison of the two pricing regimes. The solution algorithm starts with the initialization of the forward prices, which are then fixed in the forward market clearing. If the sum ϵ over the marginal of the forward quantities market clearing condition η_c (equation 4.23b) is smaller than some pre-defined solution accuracy δ, the algorithm stops.

```
Initialize $p_c^{f,bid} = 0, p_c^{f,ask} = 0$
While ($\epsilon \geq \delta$) do
    Forward market clearing
    $p_c^{f,bid} = p_c^{f,bid} + \eta_c$
    $p_c^{f,ask} = p_c^{f,bid} - \chi$
    $\epsilon = \|\eta_c\|$
end while
```

4.3.3 Results

4.3.3.1 Efficient transmission forward market

The numerical simulation with the variations presented above confirms the general insights from the previous section. Figure 4.6 shows the delta performance of nodal to zonal pricing with respect to overall welfare. A positive (negative) number indicates better performance of nodal (zonal) pricing. For every parameter variation, the minimum and average of all scenario combinations are given. The numbers should be interpreted with care, because the absolute numbers depend highly on the chosen parameter setting. The shape and behavior of the different parameter variations however, can be used to gain general insights.

The average welfare lays in between 34,120 € for nodal pricing and 25,229 € for zonal pricing. In all considered scenarios, the delta is at least zero and clearly positive in most of the cases. For the varying grid configuration, which implicitly indicates the severity

[62] The convergence of the algorithm with different starting values indicates that the equilibrium is unique.

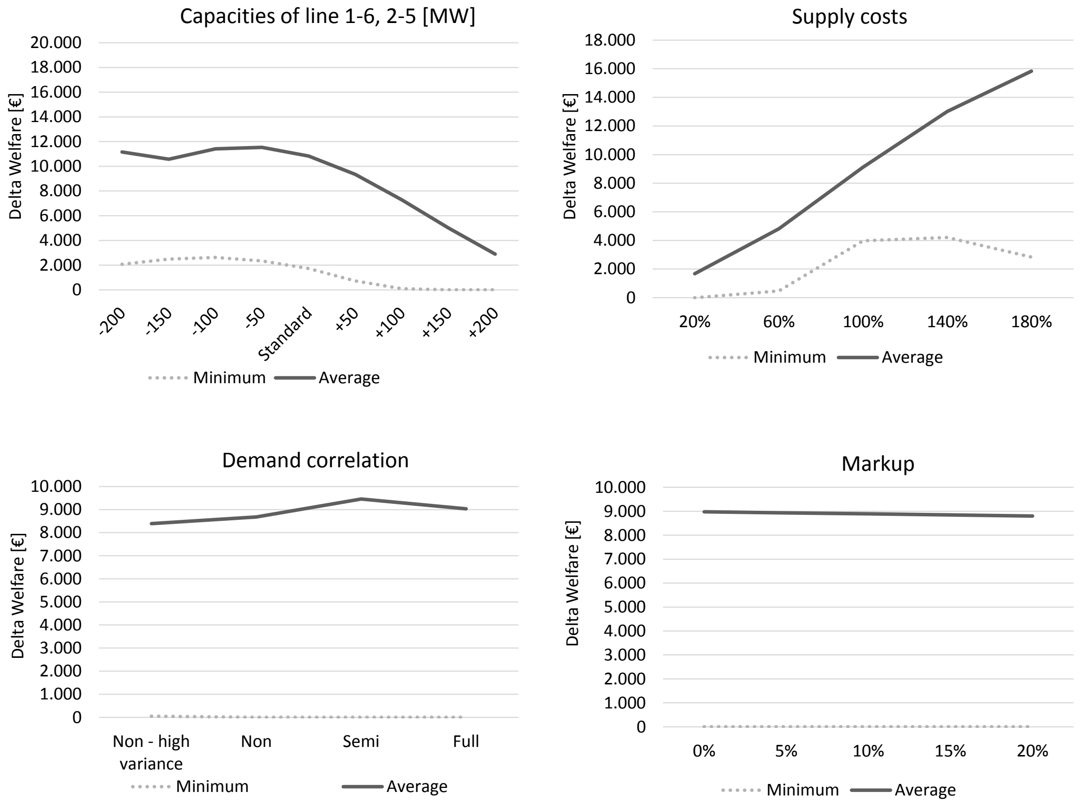


FIGURE 4.6: Delta performance of nodal to zonal pricing in the spot market

of bottlenecks, the relative performance of zonal pricing decreases with increasing grid capacities. Logically, with no bottlenecks both regimes have the same outcome. With increasing bottlenecks nodal pricing performs better. This is due to the increased costs of redispatch measures in the zonal pricing regime.[63] This effect is similar for the cost variation. If supply is cheaper, the severity of bottlenecks in terms of costs decreases due to reduced overall system costs and hence costs of redispatch. If supply becomes more costly, a decline of the delta can be observed. This effect stems from reduced quantities due to decreasing demand. If the supply costs increase further, the delta would decline to zero at the moment when the demanded quantities become zero. The delta between nodal and zonal pricing depends slightly on the demand volatility, with a higher correlation of demand increasing the performance of nodal pricing. An increasing markup has no significant effect on the relative performance of the two pricing regimes.[64]

[63] The decline in the scenario with 150 MW below the standard can be explained by the supply and demand curves as well as the power flows. For the lines 1-6 and 2-5, transmission capacity is reduced simultaneously. This impacts the welfare or more precisely the delta welfare between nodal and zonal pricing in a non-linear way, e.g., due to one node not being supplied any more in the nodal pricing regime or a different redispatch in the zonal pricing regime.

[64] Some dispatch situations occur, where total welfare is higher for zonal pricing than for nodal pricing. This is due to the calculation method of the surplus: The profit of the retailer is added ex-post, not part of the optimization and furthermore depends on the quantities and volatility. In some cases this causes (minimally) higher welfare for zonal pricing. However, this effect is due the calculation method and not dependent on the dispatch system. This partly not precise result is ignored in the further analysis.

As for the spot market, the numerical simulation for the forward market yields the same overall outcome. Nodal pricing performs better in any parameter combination supporting again the general insights from before. Figure 4.7 shows the parameter variation for the forward market. Not very surprisingly, the fundamental parameters of the spot market have the same influence in the forward market. The reason for most minima indicating $\Delta W^f = 0$ can be explained by the effect of the increased grid capacities. At some point in the variation +150 MW and +200 MW grid capacity, there is no more congestion and hence the welfare converges. The two additional parameter variations impact the costs of risk. A smaller α indicates higher costs and a higher β more impact of these costs in relation to the expectation value. The variation of α does not have much influence, which can be explained by the possibility of all market participants to find a counterpart to hedge their risks. Retailers can perfectly hedge their risks with the producers. The variation of β seems to affect the delta between the regimes. However, as for the cost factor, this is mainly due to the variation in the overall welfare measure.

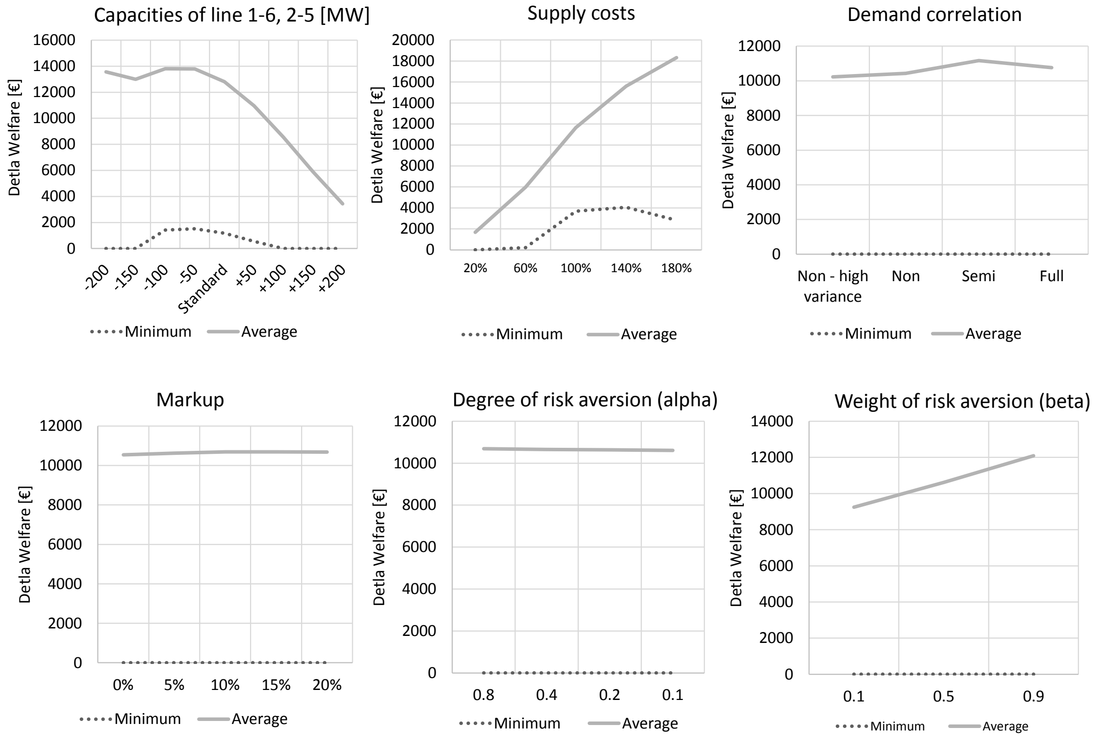


Figure 4.7: Relative performance of nodal to zonal pricing in the forward market

4.3.3.2 Inefficient transmission forward market

Now the case of an inefficient forward market for transmission is considered. Inefficiency in this context means that the forward contracts for transmission are not fully tradeable due to a positive bid-ask-spread. Hence, market participants are not able to hedge their risks up to the desired level. Obviously, an inefficient transmission forward market reduces the overall welfare in a nodal pricing regime. Unhedged risks of market participants cause additional capital requirements. The impact on welfare then depends on the nodal price volatility in the spot market which causes the risks in the first place. Furthermore, the level of the additional capital required reduces the overall welfare level. In addition, the reduced traded quantity of transmission forward contracts possibly impacts the traded quantity of energy forwards and intensifies the welfare reduction (also shown by de Maere d'Aertrycke and Smeers (2013)). If this welfare reduction is larger than the inefficiencies in the zonal spot market, a zonal performs better than a nodal pricing regime. This trade-off might appear to be simple, but it is influenced by several factors, which are fundamental for both effects. The final trade-off then depends on the structure of demand, cost differences, severity of congestion as well as the level of inefficiency in the forward transmission market and the capital requirement.

Figure 4.8 shows the number of scenarios wherein zonal pricing performs as good as or better than nodal pricing for the fundamental factors. The overall number of zonal pricing performing better is small compared to the overall number of scenarios. Despite the total percentage of zonal pricing being advantageous is small, some scenarios could still be highly relevant. Interesting are the general trends where zonal pricing performs better than nodal pricing. For the grid variation, it can be seen that zonal pricing comes closer or performs better than nodal pricing, if there is little congestion. On the one hand this is due to the overall convergence of the pricing regimes in case of no congestion. On the other hand with little congestion, redispatch costs become smaller while costs from inefficiency become more relevant. For the costs variation, small costs induce low redispatch costs and therefore make the inefficiencies of zonal pricing cheaper. For demand correlation, the results are somehow counter-intuitive at first sight, since the relative performance of zonal pricing seems to be increasing with correlation. This can again be explained by the convergence of the market outcomes if congestion is low which is the case for higher correlation.

Figure 4.9 shows the relative performance between the pricing regimes for each parameter variation relevant for the inefficiency of the forward transmission market. Increased markups reduce losses and hence the needs for hedging, leading to the straightforward result of a stabilized performance of zonal relative to nodal pricing. An increasing bid-ask-spread leads to a more inefficient market and hence, zonal pricing performance improves. Also straightforward are the results for risk aversion. With increasing risk

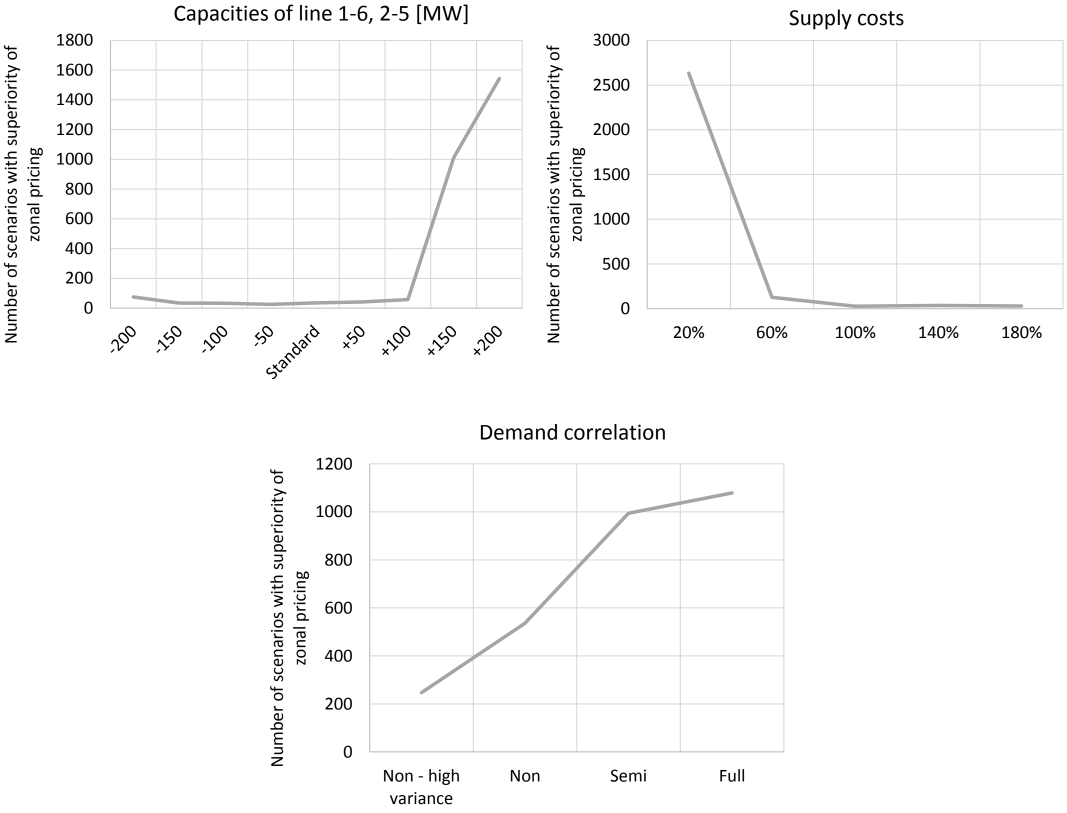


FIGURE 4.8: Number of scenarios with zonal pricing performing equally or better than nodal pricing dependent on parameter variation

aversion and increased weight put on this risk aversion, more hedging is required and it is more expensive in terms of welfare not to be able to hedge.[65]

Even a small increase of the bid-ask-spread reduces traded quantities drastically.[66] Furthermore, the impact of an inefficient transmission forward market on the traded quantity of the energy forward can be seen in Figure 4.10. A reduced volume of forward transmission trades induces a lower quantity of energy forward trades. Retailers try to hedge their local risk, for which they have to buy an energy forward at the hub and a transmission forward. If transmission forwards become less attractive due to an increased bid-ask-spread, they also reduce the number of energy forwards for their hedging.

While the reduction of volumes is drastic when introducing a bid-ask-spread, the impact on prices is lower as shown in Table 4.2.[67] Prices increase slightly with decreasing efficiency. Of course, this stems partly from averaging the values. Simulations with higher risk aversion induce higher price reductions, e.g., for $\alpha = 0.1, \beta = 0.9$ an energy

[65] The slight peak of $\beta = 0.5$ can again be explained by non-linearities caused by the underlying fundamental values.

[66] The quantities in the liquid case are significantly higher as in de Maere d'Aertrycke and Smeers (2013) which is due to a high trading activity of the TSO, which is assumed to be risk neutral.

[67] The negative value for FTR4 from node 4 to node 6 indicates that with a higher bid-ask-spread the direction of the FTR changes.

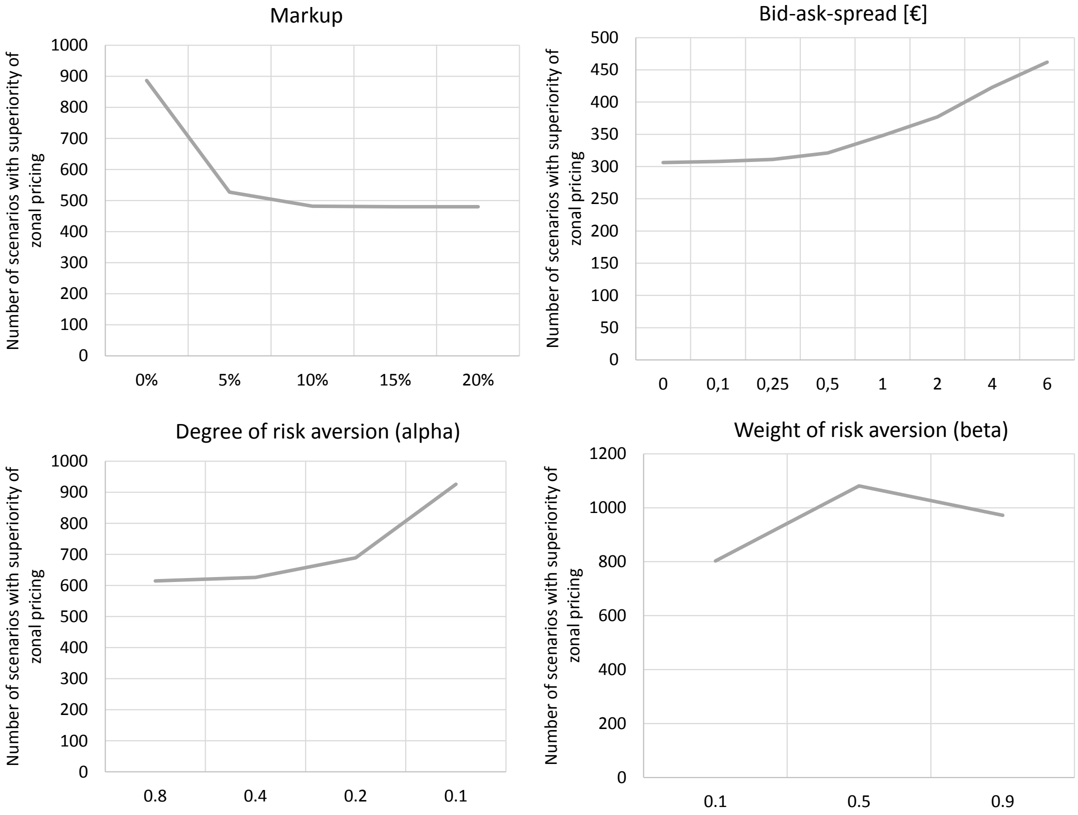


FIGURE 4.9: Number of scenarios with zonal pricing performing equal or better than nodal pricing dependent on parameter variation

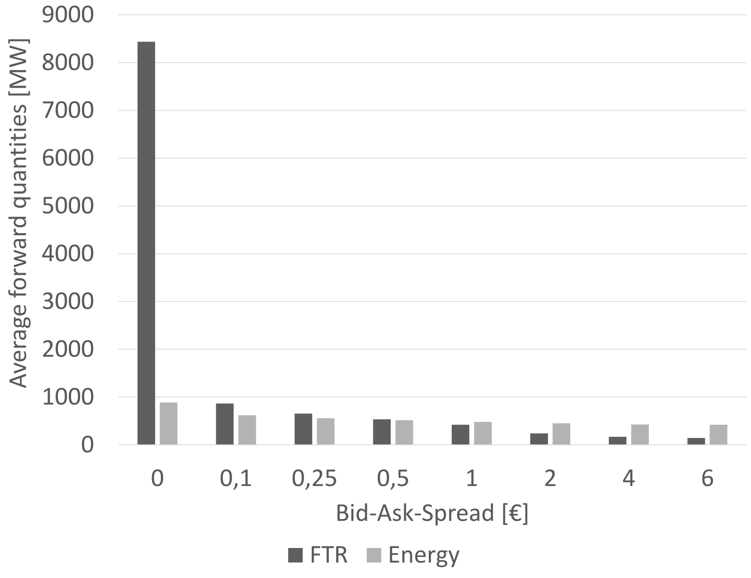


FIGURE 4.10: Average forward quantities of nodal pricing

price difference of 5 occurs between the efficient transmission market and a bid-ask-spread of 4. No shifting from one FTR to another seems to take place when looking at the prices. The reason for this however is that the bid-ask-spread is equal for all FTR, keeping the general relationships stable.

	Bid-Ask-Spread							
	0,0	**0,1**	**0,25**	**0,5**	**1,0**	**2,0**	**4,0**	**6,0**
Energy (6)	44,4	44,4	44,4	44,5	44,5	44,7	44,9	45,1
FTR1 (6-1)	20,9	21,0	21,1	21,2	21,3	21,7	22,1	22,5
FTR2 (6-2)	16,7	16,8	16,8	17,0	17,1	17,5	18,2	18,8
FTR3 (6-3)	19,1	19,2	19,3	19,4	19,6	19,9	20,6	21,1
FTR4 (4-6)	1,8	1,8	1,7	1,6	1,4	1,0	0,4	-0,1
FTR5 (6-5)	4,1	4,1	4,2	4,3	4,4	4,7	5,4	5,9

Table 4.2: Average prices for forward contracts with nodal pricing [€]

4.4 Conclusions

The literature has shown the theoretical superiority of nodal pricing compared to zonal pricing in efficient markets. Zonal pricing is inherently inefficient due to hidden scarcities of transmission constraints. Empirical work, however, showed that forward markets for financial transmission rights in nodal pricing regimes might lack efficiency impacting the performance of nodal compared to zonal pricing.

In this paper, a zonal and nodal pricing regime were compared and the impacts of an inefficient transmission forward market were analyzed. The general effects have been shown in a simple two node model. The conclusions for efficient markets were confirmed. The trade-off between an inefficient transmission forward market (in a nodal pricing regime) and the inherent inefficiencies of redispatch (in a zonal pricing regime) have been formalized. Comparative statics were performed with a model incorporating more nodes, loop flows as well as energy and transmission forwards. For this, the nodal pricing spot and forward market model by de Maere d'Aertrycke and Smeers (2013) was extended by a zonal pricing approach and a producer-only redispatch. Furthermore, the volume constraint reducing efficiency was replaced by a formulation via a bid-ask-spread.

The relative performance of the pricing regimes has been tested for a wide range of scenarios with varying demand volatility, supply costs, grid configurations, markups and risk aversion. The results for the spot market showed that nodal pricing is always performing at least as good as zonal pricing (and better in nearly all considered cases). This holds also for the case of an efficient forward market, regardless of the parameter setting. Inefficiencies in the forward transmission market, in terms of a positive bid-ask-spread and risk aversion of market participants, lead to situations wherein zonal pricing outperforms nodal pricing. Given all considered parameter variation this happens only in a relatively small number of cases. Nevertheless, this matters if these cases are the most relevant ones. It seems plausible that each pricing regime performs better, if the respective weaknesses, i.e., the inefficiency of the forward transmission market or the inherently inefficient redispatch, are highly relevant: A nodal pricing regime performs better, if congestion within a zone is severe and costly. In turn, a zonal pricing regime performs better, if the bid-ask-spread and the risk aversion are high.

The results imply that the trade-off between the respective weaknesses of the pricing regimes should be considered carefully. In larger electricity systems such as the European one, some sort of in-between solution might be favorable, i.e., by properly defining bidding zones by considering the respective inefficiencies. However, other factors such as market adaptation to newly defined zones and transaction costs have to play a role within such considerations.

Further research should clarify whether or not the findings of the rather simple numerical setting can be transferred to a more complex one. In addition, the effects of strategic behavior, different preferences, portfolio effects or uncertainty are worthwhile considering.

Regulation of non-marketed outputs and substitutable inputs

5.1 Introduction

Numerous goods and services are provided by regulated firms with a monopolistic status. For instance, uninterrupted electricity transmission services - being a textbook example of a natural monopoly - are usually provided by a single firm. Currently, an increasing deployment of renewable energy sources leads to substantially changing requirements to secure an uninterrupted electricity transmission, while multiple substitutable measures may exist to cope with it, such as grid expansion or sophisticated grid operation.[68] Due to the fact that an uninterrupted electricity transmission is crucial for society, the regulator will be well aware of whether or not it has been provided *effectively*.[69] In contrast, however, electricity systems are highly complex, such that interdependent activity levels as well as related cost figures are hard to assess. Hence, it may be difficult for the regulator to judge the *efficiency* of the firm's underlying measures. Technically speaking, this situation may be seen as a production process involving multiple substitutable inputs, incorporating two adverse selection problems: First, the regulator may have a hard time estimating the necessary overall level of the firm's activity, determined by the marginal rate of technical substitution (MRTS), i.e., the isoquant function describing the relation of inputs needed to produce the requested output. Second, the regulator may have difficulties verifying the unit costs of one or multiple inputs. This multi-dimensional asymmetric information increases the complexity of finding an adequate regulation.

In theory as well as in practice, problems of information asymmetry between the regulator and the firm have been tackled by different forms of regulation. Typical approaches in regulatory practice range from cost-based regulation to widely applied incentive regulation (discussed, e.g., in Joskow (2014)), or a linear combination of those two extremes (e.g., Schmalensee (1989)). For instance, the German regulator offers one single contract

[68] The German Transmission System Operators estimate the necessary investments into grid reinforcements and expansion to be around 22 bn. € for the period 2013-2022 (Netzentwicklungsplan, 2013), which doubles the annual figures for 2012 and quadruples the value for 2006 (Bundesnetzagentur, 2013c).

[69] For instance, in Germany the regulator has defined five observable, quantifiable dimensions for measuring grid quality.

to electricity transmission firms, dependent on grid expansion, which corresponds to a cost-based regulation of capital.[70] The academic discussion has not yet fully covered the specific multi-dimensional problems of asymmetric information regarding the level and mix of inputs, but more recent theoretical approaches suggest that the best theoretical solution consists of the regulator offering the firm a *menu* of contracts, such that the firm reveals her private information (e.g., Laffont and Tirole (1993)). Even though the dichotomy between such Bayesian models of regulation (which tend to dominate the academic discussion) and simpler non-Bayesian models (which are closer to regulatory practice) is well perceived, corresponding explanations are rather vague. For instance, as Armstrong and Sappington (2007) note, "[...] *regulatory plans that encompass options are 'complicated', and therefore prohibitively costly to implement*".

The goal of this paper is twofold: First, to identify and investigate the optimal Bayesian regulation for the multi-dimensional problem at hand, and second, to bridge the gap between the theoretically optimal solution and simpler regimes applied in regulatory practice.

To derive an optimal regulation strategy, we build on the theory of incentives and contract menus. It is well known that in a simple setting with two types of the firm, the efficient type is incentivized via a contract with first best (price) levels along with some positive rent, while the inefficient type's contract includes prices below the first best and no rent (e.g., Laffont and Tirole (1993)). This analysis has been extended to represent multiple dimensions of information asymmetry in terms of adverse selection, e.g., by Lewis and Sappington (1988b), Dana (1993), Armstrong (1999) or Aguirre and Beitia (2004). While Dana (1993) analyzes a multi-product environment, Lewis and Sappington (1988b), Armstrong (1999) and Aguirre and Beitia (2004) consider two-dimensional adverse selection with only one screening variable. Specifically, the latter three derive optimal regulation strategies in a marketed-good environment (in the sense of Caillaud et al. (1988)) with unknown cost and demand functions. In our paper, unlike Lewis and Sappington (1988b) and Armstrong (1999), we consider shadow costs of public funding instead of distributional welfare preferences. Despite technical differences, this is largely in line with the analysis of Aguirre and Beitia (2004).[71] However, in contrast to all these papers, we solve the two-dimensional adverse selection problem for a non-marketed good environment and a production process that involves two substitutable inputs with an uncertain isoquant and input factor costs.[72]

[70] In Germany, transmission system operators formulate a network expansion plan for which they get an allowed investment. In line with economic theory, the chosen levels may be suspected to be inefficiently high (see Footnote 1 for related cost figures). This regulation corresponds to a cost-based regulation for the input factor grid expansion, while neglecting any other possible input, such as better operational measures. Obviously, this triggers some sort of Averch-Johnson-effect and leads to suboptimal distortions of the input levels.

[71] Aguirre and Beitia (2004) show the difference between shadow costs of public funding and distributional welfare preferences based on a model with continuous probability distribution, while we assume a discrete distribution.

[72] Noticeably, with the (discrete) two-dimensional adverse selection problem, our problem setting is technically closest to the model discussed by Armstrong (1999).

For the novel setting of multi-dimensional inputs and a non-marketed output, we are able to confirm the general insights from the above literature. We find that expected social welfare necessarily includes positive rents for some types of the firm, such that the first best solution cannot be achieved. While the efficient type is always set to first best input levels, the other contracts' (observable) input levels are distorted upwards.[73] Separation of at least three types is always possible, while bunching of two types may be unavoidable in case of a very asymmetric distribution of costs or very flat isoquants.

We compare the obtained optimal Bayesian regulation to the results of a non-Bayesian regulation that we obtain by restricting our regulation problem to one single contract. We find that despite the general inferiority a non-Bayesian cost-based regulatory regime may indeed be close to the optimal Bayesian solution for specific circumstances. This especially holds true if the overall input level probably needs to be high, and shadow costs of public funding are large. Considering current circumstances observed in the electricity sector, i.e., substantial changes in the supply structure and ongoing intense discussions about grid tariffs, these conditions may indeed prevail.

The paper is organized as follows: Section 5.2 introduces the model, Section 5.3 presents the optimal regulation strategy, Section 5.4 compares the optimal regulation to simpler regimes, and Section 5.5 concludes.

5.2 The model

Consider a single firm that is controlled by a regulator. The firm uses two inputs to provide an output in terms of a good or service level q that is requested by the regulator. The regulator's choice of q could, for instance, result from counterbalancing the economic value of the provided with the related social costs. For simplicity, however, we assume q to be invariant throughout the paper. Although this assumption might seem restrictive at first sight, it may indeed fit a number of relevant cases very well. For instance, due to the very high societal value of uninterrupted electricity transmission, changes in costs will hardly affect the desired level of the transmission service quality q.

In our model, probability μ (respectively $1 - \mu$) leads to a low (high) aggregated input that is necessary to reach the same requested output q. This could, e.g., be an exogenous shock induced by the increased deployment of renewable energies, triggering a changing spatial distribution of supply and hence impacting the necessary overall activity level in the grid sector to achieve secure electricity transmission. From the firm's perspective, an

[73] Upwards distorted observable input levels coincide with upwards distorted prices for the inefficient type as shown in Laffont and Tirole (1993). They also agree with the results in a setting with unknown cost and demand functions as long as shadow costs of public funding are considered (Aguirre and Beitia (2004)). Noticeably, the case of prices below marginal costs, as found in Lewis and Sappington (1988b) and Armstrong (1999), is mainly triggered by using a distributive social welfare function instead of shadow costs of public funding.

output level q can be provided by means of two different inputs, one of which is observable (x) and one non-observable (y) by the regulator. Stressing again our introductory example of electricity transmission services, x could be the level of grid expansion that is easily observable by the regulator, even by people unfamiliar with the details of electricity transmission. Utilization and measures of sophisticated grid operation, especially as a partial substitute to grid expansion, however, are hardly observable. The tradeoff between those two inputs needed to reach output q is commonly described by a production function $q = f(x, y)$ which can be illustrated by means of isoquants. We assume smooth and decreasing marginal returns of both inputs, such that the isoquants are downward sloping, convex and differentiable. Noticeably, two different isoquants can never cross. An example fulfilling these requirements is a Cobb-Douglas-type production function. The inverse production function $g(q, x)$ reflects the necessary level of the non-observable input y needed to reach output q, given a level of x. We will mostly use this inverse function hereafter. Due to the exogenous shock leading to a low (l) or high (h) aggregated input necessary for the envisaged output level q, the inverse function takes one of two possible functional forms, i.e., $g_i(q, x)$, with $i \in [l, h]$ and $g_l(q, x) < g_h(q, x)$.

The optimal rate of substitution between the two inputs minimizing total costs for reaching the requested output depends on the cost functions of the inputs. We consider the cost function $c^x(x)$ of the the observable input to be fixed and common knowledge, while the cost function of the non-observable input $c^y_j(y)$ is subject to a nature draw, which leads with probability ν (respectively $1 - \nu$) to a low (high) cost function (i.e., $j \in [l, h]$). For simplicity, we assume constant factor costs of both inputs, i.e., $c^x(x) = c^x$ and $c^y_j(y) = c^y_j$. The realization of c^y_j influences the isocost line of the two inputs and hence, the optimal rate of substitution.[74] Hence, depending on the two random draws for the isoquant and the costs of the non-observable input, there are four possible first best bundles of inputs, which we denote by $\{x^{fb}_{ll}, y^{fb}_{ll}\}$, $\{x^{fb}_{lh}, y^{fb}_{lh}\}$, $\{x^{fb}_{hl}, y^{fb}_{hl}\}$ and $\{x^{fb}_{hh}, y^{fb}_{hh}\}$. As a last precondition, we require the expansion path, i.e., the curve connecting the optimal input combinations of the different isoquants, to be pointing rightwards as the necessary aggregated input increases.[75] In terms of the first best input levels, this requires $x^{fb}_{ll} > x^{fb}_{hl}$ and $x^{fb}_{lh} > x^{fb}_{hh}$, which again holds true for a wide range of possible production function specifications, including the above mentioned Cobb-Douglas type.

Under optimal Bayesian regulation, the goal of the regulator is to incentivize the firm via a suitable contract framework to choose the welfare-optimizing bundle of inputs, which we will derive based on classic mechanism design entailing truthful direct revelation. Contrary to the firm, the regulator cannot observe the realizations of the two random draws, although the possible realizations as well as the occurrence probabilities are

[74] As it is well known from production theory, the optimal rate of substitution is determined by equating the marginal rate of technical substitution between the factors (i.e., the slope of the isoquant) with the relative factor costs (i.e., the slope of the isocost line).

[75] For an analysis involving continuous variables, this would require the expansion path to behave like a function with a unique function value y for each x, or, in other words, an expansion path that is not bending backwards.

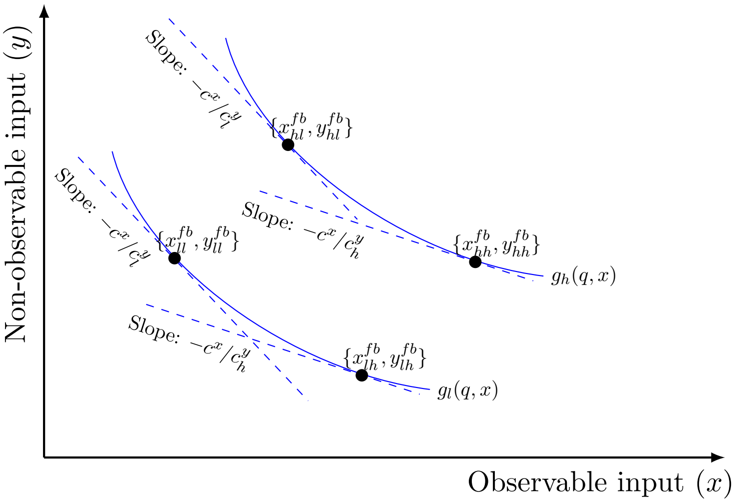

Non-observable input (y)
Observable input (x)
Slope: $-c^x/c^y_l$
$\{x^{fb}_{hl}, y^{fb}_{hl}\}$
Slope: $-c^x/c^y_h$
$\{x^{fb}_{hh}, y^{fb}_{hh}\}$
$g_h(q,x)$
Slope: $-c^x/c^y_l$
$\{x^{fb}_{ll}, y^{fb}_{ll}\}$
Slope: $-c^x/c^y_h$
$\{x^{fb}_{lh}, y^{fb}_{lh}\}$
$g_l(q,x)$

FIGURE 5.1: Problem setting with double adverse selection

common knowledge. She knows the cost function of the observable input and can observe the corresponding input level. The output is also observable and verifiable.[76] For an optimal regulation, the regulator offers the firm a menu of four contracts, each with a level of the observable input x_{ij} and a corresponding transfer T_{ij}. Naturally, the contracts can be conditioned on observable parameters only, i.e., the output as well as the amount of the observable input used. Both are enforceable by means of suitably high penalties in case the firm deviates from the requested/contracted level.

The timing – as shown in Figure 5.2 – is as follows. First, the random draws are realized and the cost function of the non-observable input and the necessary aggregated input relation (isoquant) are observed by the firm. The firm then chooses between several (in our case, four) contracts offered by the regulator. She then realizes the input levels to produce the requested output. The regulator observes one input level (x) and whether the output is as requested; if those are as agreed upon, the contract is executed and the transfer realized.

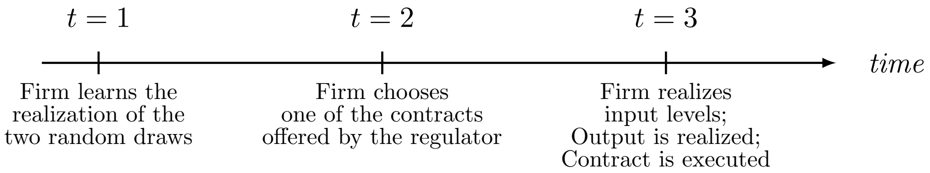

$t = 1$
$t = 2$
$t = 3$
time
Firm learns the realization of the two random draws
Firm chooses one of the contracts offered by the regulator
Firm realizes input levels; Output is realized; Contract is executed

FIGURE 5.2: Timing

The rent of the firm R_{ij} given a realization $i \in [l, h]$ and $j \in [l, h]$, results from the transfer T_{ij} minus the private cost of the firm's activities:[77]

$$R_{ij} = T_{ij} - c^x x_{ij} - c^y_j g_i(q, x_{ij}) \tag{5.1}$$

[76] Stochastic deviations due to force majeure are supposed to be detectable and excludable from the contract framework.

[77] It goes without saying here that the firm is characterized such that she tries to maximize her rent.

The regulator maximizes expected social welfare, defined as the sum of expected social utility and firm surplus, by adjusting the observables, i.e.:

$$\max_{x_{ij},T_{ij}} W = \mathbb{E}\left[\underbrace{S_q - (1+\lambda)T_{ij}}_{\text{Net social utility}} + \underbrace{(T_{ij} - c^x x_{ij} - c^y_j g_i(q, x_{ij}))}_{\text{Firm's rent } (R_{ij})}\right] \tag{5.2}$$

where S_q is the gross social utility from reaching output q, and λ denotes the shadow costs of public funding, i.e., the costs due to raising and transferring finances through public channels (for a discussion, see, e.g., Laffont and Tirole (1993)). As discussed previously, we assume q – and hence also gross social utility S_q – to be invariant and independent of the random draws, yielding[78]

$$\max_{x_{ij},T_{ij}} W = S_q - \mathbb{E}\left[\underbrace{(1+\lambda)T_{ij}}_{\text{Transfer costs}} - \underbrace{(T_{ij} - c^x x_{ij} - c^y_j g_i(q, x_{ij}))}_{\text{Firm's rent } (R_{ij})}\right] \tag{5.3}$$

As an important consequence of Equation (5.3), we see that the optimization problem of the regulator can be reformulated in terms of a cost-minimization problem, essentially stating that the desired output shall be reached at minimal expected social costs:

$$\min_{x_{ij},T_{ij}} C = \mathbb{E}\left[C_{ij}\right] = \mathbb{E}\left[\lambda \underbrace{R_{ij}}_{\text{Firm's rent}} + (1+\lambda)(\underbrace{c^x x_{ij}}_{\substack{\text{Costs of}\\ \text{observable input}}} + \underbrace{c^y_j g_i(q, x_{ij})}_{\substack{\text{Costs of}\\ \text{non-observable input}}})\right] \tag{5.4}$$

While choosing x_{ij} and T_{ij} such that social costs are minimized, the regulator is restricted by several participation and incentive constraints for the firm's rent:

$$R_{ij} \geq 0 \qquad \forall i, j \tag{5.5}$$

$$R_{ij} \geq R_{i'j'} + c^y_{j'} g_{i'}(q, x_{i'j'}) - c^y_j g_i(q, x_{i'j'}) \qquad \forall \text{ pairs } i, j \text{ and } i', j' \tag{5.6}$$

Equation (5.5) ensures that all types of firms have a non-negative profit and therefore participate.[79] In line with the revelation principle, Equation (5.6) provides the firm with the incentive to truthfully report the realized isoquant and non-observable input costs.

[78] This is the reason why q appears as a subscript here. In case of a more complex analysis involving q as a variable, S_q would be replaced by $S(q, x)$ to reflect the counterbalancing of the economic value of the provided output with the related social costs.

[79] Hence, we implicitly assume zero liability for the firm.

Written explicitly, the four participation constraints for the four possible firm types become

$$R_{ll} \geq 0 \tag{5.7a}$$
$$R_{lh} \geq 0 \tag{5.7b}$$
$$R_{hl} \geq 0 \tag{5.7c}$$
$$R_{hh} \geq 0, \tag{5.7d}$$

and the twelve incentive constraints (each of the four types might be tempted to choose a contract of one of the other three types)

$$R_{ll} \geq R_{lh} + c_h^y g_l(q, x_{lh}) - c_l^y g_l(q, x_{lh}) \tag{5.8a}$$
$$R_{ll} \geq R_{hl} + c_l^y g_h(q, x_{hl}) - c_l^y g_l(q, x_{hl}) \tag{5.8b}$$
$$R_{ll} \geq R_{hh} + c_h^y g_h(q, x_{hh}) - c_l^y g_l(q, x_{hh}) \tag{5.8c}$$
$$R_{lh} \geq R_{ll} + c_l^y g_l(q, x_{ll}) - c_h^y g_l(q, x_{ll}) \tag{5.8d}$$
$$R_{lh} \geq R_{hl} + c_l^y g_h(q, x_{hl}) - c_h^y g_l(q, x_{hl}) \tag{5.8e}$$
$$R_{lh} \geq R_{hh} + c_h^y g_h(q, x_{hh}) - c_h^y g_l(q, x_{hh}) \tag{5.8f}$$
$$R_{hl} \geq R_{ll} + c_l^y g_l(q, x_{ll}) - c_l^y g_h(q, x_{ll}) \tag{5.8g}$$
$$R_{hl} \geq R_{lh} + c_h^y g_l(q, x_{lh}) - c_l^y g_h(q, x_{lh}) \tag{5.8h}$$
$$R_{hl} \geq R_{hh} + c_h^y g_h(q, x_{hh}) - c_l^y g_h(q, x_{hh}) \tag{5.8i}$$
$$R_{hh} \geq R_{ll} + c_l^y g_l(q, x_{ll}) - c_h^y g_h(q, x_{ll}) \tag{5.8j}$$
$$R_{hh} \geq R_{lh} + c_h^y g_l(q, x_{lh}) - c_h^y g_h(q, x_{lh}) \tag{5.8k}$$
$$R_{hh} \geq R_{hl} + c_l^y g_h(q, x_{hl}) - c_h^y g_h(q, x_{hl}). \tag{5.8l}$$

5.3 Optimal regulation

5.3.1 Preparatory analysis

As a first preparatory step in the analysis we shall check whether the contract variable x is actually suitable to provide incentives to the firm to reveal her true type. To this end, we investigate whether the incentive to choose another type's contract (motivated by a potential increase in rent) regarding one of the two random draws is impacted by an adjustment of x. This is often referred to as "single crossing" conditions. For the incentive to choose another type's contract regarding the realized input cost, we find that[80]

$$\frac{\partial}{\partial x}(R_{ih}(x) - R_{il}(x)) = (c_l^y - c_h^y)g_i'(q,x) \qquad \text{for} \qquad i = l,h, \tag{5.9}$$

which is clearly greater than zero due to $c_h > c_l$ and $g_i'(q,x) < 0$. Hence, by an upwards distortion of x, we are able to reduce the incentive for the firm to choose the contract of a high cost type instead of truly revealing the realized low cost type.

Similarly, for the incentive to choose a contract for an isoquant different from the realized one, we find that

$$\frac{\partial}{\partial x}(R_{hj}(x) - R_{lj}(x)) = c_j^y(g_l'(q,x) - g_h'(q,x)) \qquad \text{for} \qquad j = l,h \tag{5.10}$$

which is greater than zero as long as $g_h'(q,x) < g_l'(q,x)$. Recalling from Section 5.2 that we have assumed rightwards pointing expansion paths (a property exhibited by a wide range of possible production function specifications, including the Cobb-Douglas type), this condition will always hold true. Hence, upwards distorting x will provide a possibility to reduce the incentive for the firm to choose the contract with a high isoquant instead of truly revealing the realized low isoquant.

The effect of changing incentives following a distortion of x helps us to derive a first characterization of the optimal solution of our regulatory problem. In fact, in order to comply with the incentive constraints (5.8a)-(5.8l) (which need to be fulfilled for the optimal solution anyway), input levels x_{ij} need to follow a certain ordering. Note that for each pair of types there are two relevant incentive constraints (e.g., Equations (5.8a) and (5.8d) for the types ll and lh). Adding those and using the above single crossing conditions, the necessary ordering can be obtained as follows:[81]

$$x_{ll} \leq x_{lh} \leq x_{hh} \tag{5.11}$$

$$x_{ll} \leq x_{hl} \leq x_{hh} \tag{5.12}$$

Moreover, from the incentive constraints (5.8a) and (5.8i) it follows that only the participation constraints (5.7b) and (5.7d) (i.e., limited liability of the lh and the hh-type) remain relevant for further analyses. In contrast, the other two participation constraints (those of the low-cost types) are implicitly fulfilled if these two incentive constraints hold.

So far unclear from the above analysis, however, is the ordering of the intermediate cases x_{lh} and x_{hl}, which depends on whether the term $R_{hl}(x) - R_{lh}(x)$ is increasing or

[80] Here and in the following, a prime denotes derivation with respect to x.

[81] For instance, adding Equations (5.8a) and (5.8d) yields $(c_l^y - c_h^y)g_l'(q,x_{lh}) \geq (c_l^y - c_h^y)g_l'(q,x_{ll})$, which, together with (5.9), implies that $x_{lh} \geq x_{ll}$.

decreasing in x. Differentiating with respect to x yields

$$\frac{\partial}{\partial x}(R_{hl}(x) - R_{lh}(x)) = (c_h^y g_l'(q,x) - c_l^y g_h'(q,x)) \tag{5.13}$$

which is increasing in x as long as

$$\frac{c_h^y}{c_l^y} < \frac{g_h(q,x)}{g_l(q,x)}, \tag{5.14}$$

and decreasing in x otherwise. Together with incentive constraints (5.8e) and (5.8h) we infer that if the cost variation is small compared to the isoquant variation, then $x_{lh} \leq x_{hl}$. If the aggregated input level variation is small compared to the cost variation, then $x_{lh} \geq x_{hl}$. For an intuition, recall Figure 5.1. If the aggregated input level variation and hence the distance between the isoquants is large, x_{hl}^{fb} is larger than x_{lh}^{fb}. If the cost variation, and hence, the vertical distance between the corresponding first best solutions is large, x_{lh}^{fb} is larger than x_{hl}^{fb}.

The results of our preparatory analysis are summarized in the following two Lemmas.

Lemma 5.1. *Limited liability is only an issue for the high-cost types. Hence, the only relevant participation constraints are* (5.7b) *and* (5.7d)*, whereas* (5.7a) *and* (5.7c) *are implicitly fulfilled.*

Lemma 5.2. *In order to reach incentive compatibility, input levels x_{ij} must be ordered as follows:*

(A) If the cost variation is small compared to the isoquant variation, then $R_{hl}(x) - R_{lh}(x)$ is increasing in x and requires

$$x_{ll} \leq x_{lh} \leq x_{hl} \leq x_{hh}. \tag{5.15}$$

(B) If the cost variation is large compared to the isoquant variation, then $R_{hl}(x) - R_{lh}(x)$ is decreasing in x and requires

$$x_{ll} \leq x_{hl} \leq x_{lh} \leq x_{hh}. \tag{5.16}$$

5.3.2 Full information benchmark

If the regulator had no information deficit, she would observe the realized isoquant as well as the realized isocost line. Differentiating all possible realizations of the social cost function C_{ij} with respect to the observable input levels x_{ij} shows that all of them are single-peaked with a unique minimum at $g_i^{'}(q, x_{ij}) = -\frac{c^x}{c_j^y}$, which is necessarily realized at $x_{ij} = x_{ij}^{fb}$. The regulator would easily derive the first best levels of inputs to supply the requested output at minimal social costs, i.e., $\{x_{ij}^{fb}, y_{ij}^{fb}\}$, by equating the known realized marginal rate of technical substitution of the inputs with the realized isocost line. Moreover, she would be able to enforce the implementation of the first best due to the full observability. The corresponding optimal transfers would be $T_{ij}^{fb} = c^x x_{ij}^{fb} + c_j^y y_{ij}^{fb}$, leaving all types of the firm with zero rent. In the case of full information, social costs amount to $C_{ij}^{fb} = (1+\lambda)T_{ij}^{fb} = (1+\lambda)(c^x x_{ij}^{fb} + c_j^y y_{ij}^{fb})$, corresponding to the welfare-optimizing first best solution that could thus be obtained.

5.3.3 Asymmetric information

In the case of asymmetric information, the only two observables for the regulator are the output q and the observable input x. In addition, she can choose an appropriate level of transfer payment T. As q is invariable and observable, its implementation can be enforced by means of suitably high penalties in case the firm deviates. Hence, x and T are the two variables the regulator will condition her contracts on. The general idea for the regulator's optimal regulation strategy is to offer a menu of contracts with optimized variables $\{x_{ij}^*, T_{ij}^*\}$, such that expected social costs are minimized (as stated in Equation (5.4)), and participation (Equation (5.5)) and incentive constraints (Equation 5.6) fulfilled. Hence, we restrict our attention to incentive compatible contracts ensuring that the firm always reveals her true type. Under these conditions, the revelation principle requires that the solution found (if any) is a Bayesian-Nash equilibrium (Myerson (1979), Laffont and Martimort (2002)).

5.3.3.1 One-dimensional asymmetric information

We shall first investigate a simplified problem with one-dimensional asymmetric information only, i.e., isoquant *or* cost uncertainty. Eliminating the isoquant uncertainty (by setting $\mu = 0$, $\mu = 1$ or $g_{lj} = g_{hj}$), we are left with two constraints binding: the participation constraint of the high cost type (5.7b or 5.7d) and the incentive constraint from the low to the high cost type (5.8a or 5.8i). This leads to the simplified cost function:

$$\begin{aligned} C = & \nu \left[\lambda \left(g_i(x_{ih})(c_h^y - c_l^y)\right) + (1+\lambda)\left(c^x x_{il} + c_l^y g_i(x_{il})\right)\right] \\ & + (1-\nu)\left[(1+\lambda)\left(c^x x_{ih} + c_h^y g_i(x_{ih})\right)\right] \end{aligned} \tag{5.17}$$

Derivating with respect to $x_{ij}, j \in l, h$ yields the following first order conditions:

$$\frac{\partial C}{\partial x_{il}} = 0 \Leftrightarrow g_i'(x_{il}^*) = -\frac{c^x}{c_l^y}, \tag{5.18}$$

$$\frac{\partial C}{\partial x_{ih}} = 0 \Leftrightarrow \underbrace{\nu\lambda(c_h^y - c_l^y)g_i'(x_{ih}^*)}_{<0} + \underbrace{(1-\nu)(1+\lambda)(c^x + c_h^y g_i'(x_{ih}^*))}_{\substack{=0 \text{ for } x_{ih}=x_{ih}^{fb} \\ <0 \text{ for } x_{ih}<x_{ih}^{fb} \\ >0 \text{ for } x_{ih}>x_{ih}^{fb}}} = 0 \tag{5.19}$$

Similarly, in case of no cost uncertainty, the observable input levels of the low isoquant types are first best, whereas the high isoquant types are distorted upwards:[82]

Lemma 5.3. *In case of asymmetric information about either costs or isoquants, the respective l-type is set to first best, while the h-type is distorted upwards compared to its first best.*

Note that the result of an adverse selection problem with one-dimensional information asymmetry on costs is well-known from the literature (e.g., Baron and Myerson (1982) or Sappington (1983)). Also note that the results concerning isoquant uncertainty are strikingly different compared to the one-dimensional demand uncertainty (which essentially corresponds to the isoquant in our setting) studied by Lewis and Sappington (1988a) or Armstrong (1999). In contrast to our model – due to neglecting shadow costs of public funding – they find that the first best can be achieved in the one-dimensional case of demand uncertainty.

5.3.3.2 Two-dimensional asymmetric information

Solving the full optimal regulation problem requires minimization of social costs, subject to all imposed four participation and twelve incentive constraints. Due to the large number of constraints, we approach the optimization by solving a relaxed problem where only a subset of the constraints is considered. To this end, we need to come up with an educated guess about the binding constraints in the optimum. If we can later show that the remaining constraints are fulfilled at the solution of the relaxed problem, we will have obtained the solution of the full problem.

We already know from Lemma 5.1 that the participation constraints of the high-cost types are the only relevant ones. Furthermore, it generally seems to be a good approach to assume the “upwards” incentive constraints, i.e., from low to high isoquant, and from low to high costs, to be binding. Moreover, it seems plausible to assume binding incentive constraints from the most efficient to an intermediate type (i.e., lh or hl), and from an intermediate type to the least efficient type. If we consider the isoquant variation

[82] Due to the obvious symmetry of the problem, we omit the detailed calculation here.

more relevant than the cost variation, assuming the incentive constraints according to the ordering shown in Lemma 5.2, Case *(A)*, to be binding appears to be the most educated guess we can come up with.[83] Hence, we assume that incentive constraints (5.8a) ($ll \to lh$), (5.8e) ($lh \to hl$), and (5.8i) ($hl \to hh$) are fulfilled with equality. In addition, we assume the participation constraint of the hh-type to bind since this is the only type remaining that is not attracted by any other type. Figure 5.3 illustrates with arrows the binding incentive constraints, such that the former type is not attracted by the latter type-contract. Diamonds mark the binding participation constraints.

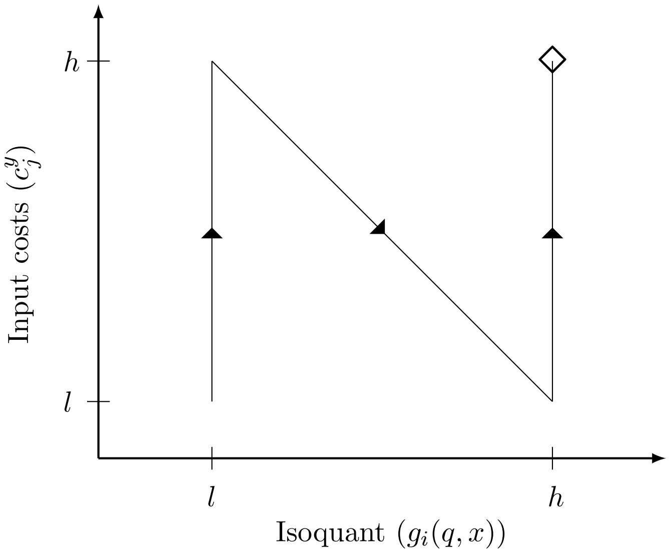


FIGURE 5.3: Constraints considered binding for Case *(A)*

We find that this set of assumptions does indeed lead us to the optimal regulation strategy. The results are summarized in the following Proposition 5.4.

Proposition 5.4. *For Case (A),*

(i) Optimal regulation is achieved under the following set of observable input levels:

$$x_{ll}^* = x_{ll}^{fb} \tag{5.20}$$

$$x_{lh}^* \geq x_{lh}^{fb} \tag{5.21}$$

$$x_{hl}^* \geq x_{hl}^{fb} \tag{5.22}$$

$$x_{hh}^* \geq x_{hh}^{fb}, \tag{5.23}$$

while respecting $x_{ll}^* < x_{lh}^* \leq x_{hl}^* \leq x_{hh}^*$.

(ii) The most efficient (ll) type can always be separated. Moreover, separation of at least three types is always possible, while bunching of the lh and hl types is unavoidable in case of $\nu \to 1$. *The hl and hh types may need to be bunched in case of* $g_l'(q,x) \to 0$ *together with* c_l^y *being large.*

[83] The ordering and solution of Case *(B)* is reversed, but similar. The corresponding discussion can be found in the appendix.

Proof. See Appendix. □

Corollary 5.5. *For $\lambda = 0$, the optimal solution is first best. All input levels amount to $x^*_{ij} = x^{fb}_{ij}$, and expected social costs to $C = C^{fb}$.*

Proof. Follows immediately from the solution of Case *(A)* when setting $\lambda = 0$. □

According to Corollary 5.5, with no shadow costs of public funding, all input levels x^*_{ij} are first best. The regulator optimizes overall welfare, but has no preference regarding the distribution of social surplus. Hence, she can give the firm an arbitrarily high budget at no social costs, and the firm maximizes her rent by setting efficient input levels. In this case, the maximization of the firm and the maximization of social welfare coincide, i.e., there is no problem of aligning the activities of the firm with social interests. Of course, larger parts of the welfare are then given to the firm.

For the general case of $\lambda > 0$, observable input levels of all types besides the ll-one are distorted above first best levels, leading to a second best solution only. Naturally, the overall level of inefficiency increases in λ, but also for decreasing μ and ν (i.e., when there is a high probability for "costly" outcomes of the random draws) as well as for $c^y_h - c^y_l$ and $g_h(q,x) - g_l(q,x)$ getting large. In contrast, however, the less significant the cost variation becomes compared to the isoquant variation, the more efficient the solution will be.

Due to keeping the most efficient (ll) type at first best level combined with the ordering according to Lemma 5.2, the type can always be separated in the contract framework. Moreover, we find that at least three types can always be separated, while bunching of two types may be unavoidable in case of vanishing isoquant or cost uncertainties, or if the isoquant variation becomes extremely large. As a last remark, it is worth mentioning that the ordering of rents is (and must be) as depicted in Figure 5.3, i.e., $0 = R^*_{hh} < R^*_{hl} < R^*_{lh} < R^*_{ll}$.

The results for Case *(B)* are symmetric but structurally identical to Case *(A)*, i.e., the ll-type is incentivized to first best input levels while the other types show upwards distortions of x_{ij}. However, roles of isoquants and costs are interchanged, reflected in the inverse occurrence of the terms $g_i \leftrightarrow c^y_j$ and $\mu \longleftrightarrow \nu$. At the same time, as imposed by Lemma 5.2, Case *(B)*, the sequence of the "intermediate" types is now $hl \to lh$. Hence, the ordering of observable input levels x_{lh} and x_{hl} as well as rents R_{lh} and R_{hl} need to be reversed to obtain an optimal regulatory contract framework.[84]

[84] See the appendix for a detailed discussion and the corresponding proposition and proof.

5.4 Comparing the optimal regulation to simpler regimes

In contrast to the optimal Bayesian menu of contracts studied in the previous section, regulatory authorities often apply alternative, simpler approaches. In fact, in the case of electricity transmission grids, it appears that they mostly offer a non-Bayesian, i.e., *single*, contract, while the application of Bayesian contracts in terms of *menus* of contracts, has been very rare.[85] For instance, regulatory practice in Germany is such that TSOs formulate a grid expansion plan, which is then reviewed and approved by the regulator. For the approved measures, the TSOs get their costs reimbursed. This corresponds to a cost-based regulation for the input factor grid expansion, while neglecting any other possible input, such as better operational measures. Meanwhile, driven, e.g., by social acceptance issues, the regulator is expected to limit the approval of extensive grid expansion to some ”reasonable” level.

Transferring such a simple non-Bayesian approach into our model, we need to limit the set of regulatory choice variables to one single contract with contract variables $\bar{x}$ and $\bar{T}$, such that the objective function of the regulator (in contrast to Equation (5.4) as in the case of optimal regulation) becomes:

$$\min_{\bar{x},\bar{T}} \bar{C} = \mathbb{E}\left[\lambda \underbrace{\bar{R}_{ij}}_{\text{Firm's rent}} +(1+\lambda)(\underbrace{c^x\bar{x}}_{\substack{\text{Costs of}\\ \text{observable input}}} + \underbrace{c^y_j g_i(q,\bar{x})}_{\substack{\text{Costs of}\\ \text{non-observable input}}})\right] \tag{5.24}$$

In contrast to the solution of the optimal regulation, this minimization is only subject to the participation constraints (5.5). With a sole contract and hence, only one observable input $\bar{x}$ for all types, the regulator has no possibility to separate types, which makes the incentive constraints obsolete. As before, the only participation constraint holding with equality is the one of the *hh*-type. Considering that this type gets full cost reimbursement but cannot be distinguished from the other types, it becomes clear that all other types must then necessarily receive a positive rent. The following proposition summarizes the solution of this non-Bayesian regulatory approach.[86]

Proposition 5.6. *Under a single contract cost-based regulation with quantity restriction, the optimal input level $\bar{x}^*$ represents an expected average of the first best solutions of the four possible types, adjusted by some upwards distortion in case of $\lambda > 0$. As an expected average, it lies between the extreme types' first best input levels, i.e., $x_{ll}^{fb} < \bar{x}^* < x_{hh}^{fb}$.*

[85] The system operator for England and Wales and the electric distribution companies in the UK are the only two examples for menus of contracts being applied in regulatory practice Joskow (2014).

[86] Note that the solution for a pure cost-based regulation without quantity restriction would simply reimburse the costs of the observable input. This would incentivize the firm to choose infinitely high values of x (known as the gold-plating effect). Assuming that the regulator restricts her set of choices by an upper level of $\bar{x} = x_{hh}^{fb}$ in order to limit excessive (socially costly) rents, all types would then choose this level. In contrast to this *very* simple approach, the regulatory regime considered in this section makes use of being able to use the observable input x as a contracting variable.

Proof. See Appendix. □

Proposition 5.7. *Compared to a single contract, the regulatory approach based on a menu of contracts is superior with respect to expected social welfare.*

Proof. It is easy to show that the optimal solution of the single contract is a feasible solution of the menu of contracts problem. Due to the fact that the solution for the menu of contracts, as stated in Proposition 5.4, is both optimal and different from the one in Proposition 5.6, it must necessarily be superior. □

As stated in Proposition 5.7, the solution of the single contract regime is always inferior to the one obtained with the menu of contracts. Nevertheless, the characteristics of the different regimes can be compared and deserve a closer look. We contrast the outcome of the optimal menu of contracts with the one of the single contract regime considering three aspects: input levels, cost-efficiency of the input levels, and rents of the firm.

Input levels for the different types have been characterized in Proposition 5.4 for the menu of contract, stating that all types besides the ll-one are distorted above first best levels. According to Proposition 5.7, the optimal input level for the single contract regime, $\bar{x}^*$, represents an expected average of the first best solution of the four possible types, adjusted by some upwards distortion in case of $\lambda > 0$. Hence, chosen input levels are generally different. However, $\bar{x}^*$ may get close to x^*_{hl} in case of λ being large and μ small (i.e., for a high probability of realizing a high isoquant). At the same time, it will never be as high as x^*_{hh}, due to $\bar{x}^* < x^{fb}_{hh} < x^*_{hh}$.

Cost-efficiency of the input levels is closely connected to the input levels and their deviation from the first best optimal solution. The optimal menu of contracts approaches first best cost-efficiency of the input levels for $\lambda \to 0$, as input levels then converge towards first best levels, i.e., $\{x^*_{ij}, y^*_{ij}\} \to \{x^{fb}_{ij}, y^{fb}_{ij}\}$. In contrast, cost-efficiency is poor for the single contract regime under this condition. However, first best input levels may also be reached, but only under very restrictive conditions, namely if $\lambda \to 0$ *and* the occurrence probability for one specific type is particularly large (e.g., if $\mu, \nu \to 1$). Type-specific as well as expected cost-efficiency of input levels is (only) then approaching first best optimality for both contracting frameworks. For the general case of $\lambda \geq 0$, it is clear that cost-efficiency of the input levels is inferior for the ll type in the single contract regime, while the ordering is ambiguous for all other types, depending on the optimal choice of $\bar{x}^*$ in comparison to x^*_{ij}.

Regarding **rents of the firm**, remember that they are only an issue for social welfare if there are shadow costs of public funding, i.e., if $\lambda > 0$. Then, however, the well known trade-off for rent-extraction and efficiency becomes relevant. For both contracting regimes, the rent of the inefficient hh type is set to zero. Moreover, for both regimes it holds true that $0 = R^*_{hh} < R^*_{hl} \ll R^*_{lh} < R^*_{ll}$ (respectively, $0 = \bar{R}^*_{hh} < \bar{R}^*_{hl} \ll \bar{R}^*_{lh} < \bar{R}^*_{ll}$), if isoquant variation is more relevant than cost variation. For the rent of specific types, we find that $R^*_{hl} < \bar{R}^*_{hl}$, while the ordering of other types' rent is generally ambiguous. Interestingly, however, if $g'_i(x)$ is small in the relevant range, $R^*_{ij} < \bar{R}^*_{ij}$ for all i, j.

Based on the above comparative statics, a singular interesting constellation can be identified for which the two contracting frameworks effectively approach each other.

Proposition 5.8. *For λ being large and μ small, the performance of the single contract is close to the one of the menu of contracts.*

Proof. For λ being large, $\bar{x}^*$ is distorted upwards (see Proposition 5.6), while $x^*_{hl} \approx x^{fb}_{hl}$ for μ small. Hence, in this case, $\bar{x}^* \approx x^*_{hl}$. Moreover, due to the fact that we consider Case (A) where cost uncertainty is relatively low, we know that the upwards distortion of x^*_{hh} is low (see Equation (5.29)), such that x^*_{hl} is not far from x^*_{hh}. Under these conditions, $\bar{C}^* \approx C^*$. □

Transferring Proposition 5.8 to our example of electricity transmission services and the regulation of the German TSOs, one may indeed come to the conclusion that the practically applied non-Bayesian regulatory approach could be close to the optimal second-best strategy. In fact, a high overall input level appears to be likely due to the strongly changing supply infrastructure, while ongoing intense discussions about the burden of electricity costs and grid tariffs for consumers could indicate high shadow costs of public funding. In the end, however, reasons for the chosen regulation are probably manifold, and might also include an explicit disutility of grid expansion, a commitment problem,[87] or the prohibitively high costs of implementing a 'complicated' regulatory regime (Armstrong and Sappington, 2007).

5.5 Conclusion

We considered a regulated firm providing a non-marketed output with substitutable inputs. We presented the optimal Bayesian regulation in terms of a menu of contracts when the regulator faces information asymmetries regarding the aggregated input level needed to provide the output as well as the realized optimal marginal rate of substitution between the inputs. Finally, the optimal Bayesian regulation was compared to a simpler non-Bayesian approach which appears to be closer to regulatory practice.

[87] Noticeably, a commitment problem of the regulator might impede the implementation of an incentive-based approach, which would be welfare-superior compared to a cost-based regulation. If the firm gets an unconditional payment representing the pay-off of the hh-type, i.e., $\tilde{T} = c^x x^{fb}_{hh} + c^y_h g_h(q, x^{fb}_{hh})$, she will realize first best input quantities $\{x^{fb}_{ij}, y^{fb}_{ij}\}$. In this case, the realized rent of the firm becomes $R_{ij} = c^x x^{fb}_{hh} + c^y_h g_h(q, x^{fb}_{hh}) - c^x x^{fb}_{ij} - c^y_j g_i(q, x^{fb}_{ij})$. However, due to the (observable) separation of types via the realized input x, the regulator might be tempted to adjust the regulatory contract ex-post, and hence, jeopardize the regulatory success if the firm anticipates this behavior.

We found that in the optimal Bayesian regulation, the first best solution cannot be achieved under the considered information asymmetries and shadow costs of public funding. This implies a strictly positive rent for the firm. The second best solution that we then characterized depends on the relative importance of the information asymmetries. However, the most efficient type is always set to first best, while the levels of the observable input are distorted upwards for all other types. At least three types can always be separated, while bunching of two types may be unavoidable in case of a very asymmetric distribution of costs or very flat isoquants. These results are structurally similar to the solutions for multi-dimensional adverse selection problems in the literature (e.g., Lewis and Sappington (1988b), Armstrong (1999) or Aguirre and Beitia (2004)). However, in contrast to existing results, our model explains upwards distortions of input levels rather than prices. Hence, we obtained important insights regarding the optimal mechanism design in the context of a regulated monopolistic firm producing a non-marketed good with multi-dimensional inputs.

The comparison to a single contract cost-based approach, as it is often applied in regulatory practice, showed that the menu of contracts is welfare superior. However, there are situations in which the performance of the approaches converge, namely if the overall input level probably needs to be high, and shadow costs of public funding are large. Given our motivating example of electricity transmission services and the current situation, e.g., in Germany, these circumstances may indeed prevail, possibly explaining the gap between the theoretically optimal Bayesian approach and the simpler non-Bayesian regulation applied in practice.

Lastly, we note that our general approach as well as our insights might also be applicable to other industries that show similar characteristics, such as public works or administrative services. Besides investigating such areas of application, future research could relax the limited liability assumption and hence, allow for a shut down of firms. Another expansion could allow the good to be marketed, which would trigger a demand reaction of the regulator (or consumers) and possibly lead to interesting variations of the conclusions derived in this paper.

Appendix

Proof of Proposition 1

Proof. (i) Under the constraints considered binding for Case *(A)* – as discussed and shown in Figure 5.3 – the social cost function (5.4) becomes

$$\begin{aligned}
&C = \\
&\mu\nu[\lambda\left(g_h(x_{hh})(c_h^y - c_l^y) + c_l^y g_h(x_{hl}) - c_h^y g_l(x_{hl}) + g_l(x_{lh})(c_h^y - c_l^y)\right) \\
&+ (1+\lambda)\left(c^x x_{ll} + c_l^y g_l(x_{ll})\right)] \\
&+ \mu(1-\nu)[\lambda\left(g_h(x_{hh})(c_h^y - c_l^y) + c_l^y g_h(x_{hl}) - c_h^y g_l(x_{hl})\right) \\
&+ (1+\lambda)\left(c^x x_{lh} + c_h^y g_l(x_{lh})\right)] \\
&+ (1-\mu)\nu\left[\lambda\left(g_h(x_{hh})(c_h^y - c_l^y)\right) + (1+\lambda)\left(c^x x_{hl} + c_l^y g_h(x_{hl})\right)\right] \\
&+ (1-\mu)(1-\nu)\left[(1+\lambda)\left(c^x x_{hh} + c_h^y g_h(x_{hh})\right)\right].
\end{aligned} \tag{5.25}$$

To derive the optimal observable input levels, we need to derive the above equation with respect to each of the four possible x_{ij}. Minimizing C with respect to x_{ll} yields

$$g_l'(x_{ll}^*) = -\frac{c^x}{c_l^y}, \tag{5.26}$$

which implies that $x_{ll}^* = x_{ll}^{fb}$. Derivations of C with respect to x_{lh}, x_{hl} and x_{hh} take the following forms:

$$\frac{\partial C}{\partial x_{lh}} = \underbrace{\mu\nu\lambda(c_h^y - c_l^y)g_l'(x_{lh})}_{<0} + \underbrace{\mu(1-\nu)(1+\lambda)(c^x + c_h^y g_l'(x_{lh}))}_{\substack{=0 \text{ for } x_{lh}=x_{lh}^{fb} \\ <0 \text{ for } x_{lh}<x_{lh}^{fb} \\ >0 \text{ for } x_{lh}>x_{lh}^{fb}}} \tag{5.27}$$

$$\frac{\partial C}{\partial x_{hl}} = \underbrace{\mu\lambda(c_l^y g_h'(x_{hl}) - c_h^y g_l'(x_{hl}))}_{<0} + \underbrace{(1-\mu)\nu(1+\lambda)(c^x + c_l^y g_h'(x_{hl}))}_{\substack{=0 \text{ for } x_{hl}=x_{hl}^{fb} \\ <0 \text{ for } x_{hl}<x_{hl}^{fb} \\ >0 \text{ for } x_{hl}>x_{hl}^{fb}}} \tag{5.28}$$

$$\frac{\partial C}{\partial x_{hh}} = \underbrace{(\mu + (1-\mu)\nu)\lambda g_h'(x_{hh})(c_h^y - c_l^y)}_{<0} + \underbrace{(1-\mu)(1-\nu)(1+\lambda)(c^x + c_h^y g_h'(x_{hh}))}_{\substack{=0 \text{ for } x_{hh}=x_{hh}^{fb} \\ <0 \text{ for } x_{hh}<x_{hh}^{fb} \\ >0 \text{ for } x_{hh}>x_{hh}^{fb}}}. \tag{5.29}$$

From Equation (5.27), we see that $\frac{\partial C}{\partial x_{lh}}$ is strictly smaller than 0 for $x_{lh} = x_{lh}^{fb}$ and monotonically increasing in x_{lh}, which implies that $x_{lh}^* > x_{lh}^{fb}$ must always hold. The same logic applies for x_{hl}^* and x_{hh}^*.

(ii) From the fact that $x_{ll}^{fb} < x_{lh}^{fb}$ and the strict upwards distortion of all other types, it follows that the ll-type can always be separated. In order to investigate whether the types lh, hl and hh can be separated or need to be bunched, we proceed as follows: For each of the possible pairs $lh - hl$, $hl - hh$ and $lh - hh$, we check the derivative of C with respect to the former type at the optimal level of x^* of the latter type (derived from the first order condition). If the change in C is greater than 0 we can conclude that we have already surpassed the optimal level of the former type, which then must be smaller than the optimal level of the latter type. In other words, we check the level of upwards distortion for the lh, hl and hh types while considering the necessary ordering of the types according to Lemma 5.2. For the pair lh-hl, we find that x_{lh}^* may surpass x_{hl}^* in case of $\nu \to 1$, while they are otherwise clearly separated from each. For the pair hl-hh, bunching may occur for $g_l'(q, x) \to 0$ together with c_l^y being large. Furthermore, we find that lh-hh can always be separated, implying that at most two types (i.e., either lh-hl or hl-hh) may be bunched under certain parameter constellations.

Lastly, it is straightforward to check that the remaining constraints are satisfied under the obtained solution of the relaxed problem. Hence, we have indeed obtained to optimal solution for the full regulatory problem we are facing in Case *(A)*. □

Proof of Proposition 2

Proof. Written explicitly, Equation (5.24) becomes

$$\begin{aligned}\bar{C} =&\mu\nu \left[\lambda \left(c_h^y g_h(\bar{x}) - c_l^y g_l(\bar{x})\right) + (1+\lambda)\left(c^x \bar{x} + c_l^y g_l(\bar{x})\right)\right] \\ &+ \mu(1-\nu)\left[\lambda \left(c_h^y g_h(\bar{x}) - c_h^y g_l(\bar{x})\right) + (1+\lambda)\left(c^x \bar{x} + c_h^y g_l(\bar{x})\right)\right] \\ &+ (1-\mu)\nu \left[\lambda \left(c_h^y g_h(\bar{x}) - c_l^y g_h(\bar{x})\right) + (1+\lambda)\left(c^x \bar{x} + c_l^y g_h(\bar{x})\right)\right] \\ &+ (1-\mu)(1-\nu)\left[(1+\lambda)\left(c^x \bar{x} + c_h^y g_h(\bar{x})\right)\right]. \end{aligned} \tag{5.30}$$

Deriving the above with respect to $\bar{x}$ yields, after a few calculations, $\mathbb{E}(g_i'(\bar{x}^*))\mathbb{E}(c_j^y) + c^x + \lambda(c_h^y g_h'(\bar{x}^*) + c^x) = 0$. Hence, for $\lambda = 0$, $\mathbb{E}(g_h'(\bar{x}^*)) = -\frac{c^x}{\mathbb{E}(c_j^y)}$. □

Two-dimensional asymmetric information, Case *(B)*: Cost variation large compared to isoquant variation

To solve the second case following from Lemma 5.2, we need to apply a different educated guess with respect to the binding constraints. However, we apply a similar reasoning as in Case *(A)*, but take account of the fact that now, cost variation is more relevant than isoquant variation. Hence, we choose a symmetric setting and imply incentive constraints (5.8b) ($ll \to hl$), (5.8h) ($hl \to lh$) and (5.8f) ($lh \to hh$) to be binding.

Again, we assume the participation constraint of the hh-type to be binding. Figure 5.4 illustrates this setting.

After having determined the results and checked all remaining constraints, we find the setting of binding constraints as in Figure 5.4 indeed to be optimal for Case *(B)*. Results are summarized in the following Proposition 5.9.

Proposition 5.9. *For case (B),*

(i) Optimal regulation is achieved under the following set of observable input levels:

$$x_{ll}^{*} = x_{ll}^{fb} \tag{5.31}$$

$$x_{lh}^{*} \geq x_{lh}^{fb} \tag{5.32}$$

$$x_{hl}^{*} \geq x_{hl}^{fb} \tag{5.33}$$

$$x_{hh}^{*} \geq x_{hh}^{fb}, \tag{5.34}$$

while respecting $x_{ll}^{*} < x_{hl}^{*} \leq x_{lh}^{*} \leq x_{hh}^{*}$.

(ii) The most efficient (ll) type can always be separated. Moreover, separation of at least three types is always possible, while bunching of the hl and lh types is unavoidable in case of $\mu \to 1$.*The lh and hh types may be bunched in case of* c_l^y *being small and* $g_h'(q,x)$ *large.*

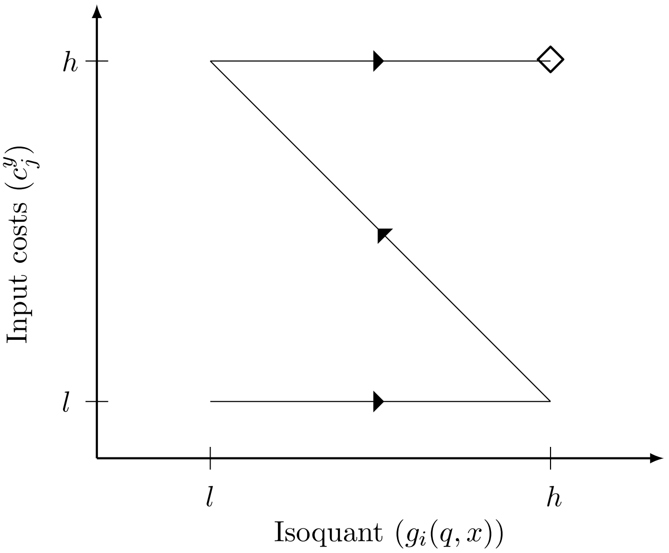


FIGURE 5.4: Constraints considered binding for Case *(B)*

Proof. (i) Under the constraints considered binding for Case *(B)* – as discussed and shown in Figure 5.4 – the social cost function (5.4) becomes

$$\begin{aligned}
&C = \\
&\mu\nu[\lambda\left(c_h^y(g_h(x_{hh}) - g_l(x_{hh})) + c_h^y g_l(x_{lh}) - c_l^y g_h(x_{lh}) + c_l^y(g_h(x_{hl}) - g_l(x_{hl}))\right) \\
&\; + (1+\lambda)\left(c^x x_{ll} + c_l^y g_l(x_{ll})\right)] \\
&\; + \mu(1-\nu)\left[\lambda\left(c_h^y(g_h(x_{hh}) - g_l(x_{hh}))\right) + (1+\lambda)\left(c^x x_{lh} + c_h^y g_l(x_{lh})\right)\right] \\
&\; + (1-\mu)\nu \\
&\left[\lambda\left(c_h^y(g_h(x_{hh}) - g_l(x_{hh}))\right) + c_h^y g_l(x_{lh}) - c_l^y g_h(x_{lh}) + (1+\lambda)\left(c^x x_{hl} + c_l^y g_h(x_{hl})\right)\right] \\
&\; + (1-\mu)(1-\nu)\left[(1+\lambda)\left(c^x x_{hh} + c_h^y g_h(x_{hh})\right)\right]. \qquad (5.35)
\end{aligned}$$

Minimizing C with respect to x_{ll} yields

$$g_l'(x_{ll}^*) = -\frac{c^x}{c_l^y}, \qquad (5.36)$$

which implies that $x_{ll}^* = x_{ll}^{fb}$. Derivation of C with respect to x_{lh}, x_{hl} and x_{hh} yields:

$$\frac{\partial C}{\partial x_{lh}} = \underbrace{\mu\lambda(c_h^y g_l'(x_{lh}) - c_l^y g_h'(x_{lh})}_{<0} + \underbrace{\mu(1-\nu)(1+\lambda)(c^x + c_h^y g_l'(x_{lh}))}_{\substack{=0 \text{ for } x_{lh}=x_{lh}^{fb} \\ <0 \text{ for } x_{lh}<x_{lh}^{fb} \\ >0 \text{ for } x_{lh}>x_{lh}^{fb}}} \qquad (5.37)$$

$$\frac{\partial C}{\partial x_{hl}} = \underbrace{\mu\nu\lambda(c_l^y g_h'(x_{hl}) - c_l^y g_l'(x_{hl}))}_{<0} + \underbrace{(1-\mu)\nu(1+\lambda)(c^x + c_l^y g_h'(x_{hl}))}_{\substack{=0 \text{ for } x_{hl}=x_{hl}^{fb} \\ <0 \text{ for } x_{hl}<x_{hl}^{fb} \\ >0 \text{ for } x_{hl}>x_{hl}^{fb}}} \qquad (5.38)$$

$$\frac{\partial C}{\partial x_{hh}} = \underbrace{(\mu + (1-\mu)\nu)\lambda c_h^y(g_h'(x_{hh}) - g_l'(x_{hh}))}_{<0} + \underbrace{(1-\mu)(1-\nu)(1+\lambda)(c^x + c_h^y g_h'(x_{hh}))}_{\substack{=0 \text{ for } x_{hh}=x_{hh}^{fb} \\ <0 \text{ for } x_{hh}<x_{hh}^{fb} \\ >0 \text{ for } x_{hh}>x_{hh}^{fb}}}. \qquad (5.39)$$

From Equation (5.37), we see that $\frac{\partial C}{\partial x_{lh}}$ is strictly smaller than 0 for $x_{lh} = x_{lh}^{fb}$ and monotonically increasing in x_{lh}, which implies that $x_{lh}^* > x_{lh}^{fb}$ must always hold. The same logic applies for x_{hl}^* and x_{hh}^*.

(ii) From $x_{ll}^{fb} < x_{lh}^{fb}$ and the strict upwards distortion of all other types, it follows that the ll-type can always be separated. x_{hl}^* may surpass x_{lh}^* in case of $\mu \to 1$. If the low costs c_l^y are small and $g_h'(q,x)$ becomes large, lh and hh types may need to be bunched, without impacting the separation of the other types.

The remaining constraints are satisfied under the obtained solution. □

As in Case (A), the first best solution can be obtained for $\lambda = 0$, while the solution is second best and incurring an increasing level of inefficiency for increasing levels of λ. Also again, the most efficient type can always be separated, while bunching of the hl and lh types (lh and hh types) may occur for very high occurrence probability of low isoquants, or if $g_h(q, x)$ is very steep and c_l^y small.

Bibliography

ACER, March 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., May 2013. Smart Modeling of Optimal Integration of High Penetration of PV - Smooth PV. Final Report.

Adamson, S., Noe, T., Parker, G., 2010. Efficiency of financial transmission rights markets in centrally coordinated periodic auctions. Energy Economics 32, 771–778.

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

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.

Aguirre, I., Beitia, A., 2004. Regulating a monopolist with unknown demand: Costly public funds and the value of private information. Journal of Public Economic Theory 6 (5), 693–706.

Andersson, G., 2011. Power System Analysis. Eidgenössische Technische Hochschule Zürich (ETH).

Armstrong, M., 1999. Optimal regulation with unknown demand and cost functions. Journal of Economic Theory 84, 196–215.

Armstrong, M., Sappington, D., 2007. Chapter 27 recent developments in the theory of regulation. In: Armstrong, M., Porter, R. (Eds.), Handbook of Industrial Organization. Elsevier, pp. 1557–1700.

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

Baron, D. P., Myerson, R. B., 1982. Regulating a monopolist with unknown costs. Econometrica: Journal of the Econometric Society 50 (4), 911–930.

Bartholomew, E. S., Siddiqui, A. S., Marnay, C., Oren, S. S., November 2003. The new york transmission congestion contract market: Is it truly working efficiently? The Electricity Journal.

Bazaraa, M. S., Sherali, H. D., Shetty, C. M., 2006. Nonlinear Programming - Theory and Algorithms. John Wiley & Sons.

Benders, J. F., 1962. Partitioning procedures for solving mixed-variables programming problems. Numerische Mathematik 4, 238–252.

Bertsch, J., Hagspiel, S., Just, L., 2015. Congestion management in power systems - long-term modeling framework and large-scale application. EWI Working Paper 15/03.

Bessembinder, H., Lemmon, M. L., 2002. Equilibrium pricing and optimal hedging in electricity forward markets. The Journal of Finance 57 (3), 1347–1382.

Bjørndal, M., Jørnsten, K., 2001. Zonal pricing in a deregulated electricity market. The Energy Journal 22 (1), 51–73.

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

Brunekreeft, G., Neuhoff, K., Newbery, D., 2005. Electricity transmission: An overview of the current debate. Utilities Policy 13 (2), 73–93.

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.

Bundesnetzagentur, 2014. Monitoring Report 2014.

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

Bundesnetzagentur, December 2015b. Quartalsbericht zu netz- und systemsicherheitsmaßnahmen: Erstes und zweites quartal 2015.

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.

Caillaud, B., Guesnerie, R., Rey, P., Tirole, J., 1988. Government intervention in production and incentives theory: A review of recent contributions. The RAND Journal of Economics 29 (1), 1–26.

CAISO, January 2000. Response to 'nodal and zonal congestion management and the exercise of market power' in "answer of the california independent system operator corporation to motions to intervene and response comments".

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

Caramanis, M., December 1982. Investment decisions and long-term planning under electricity spot pricing. IEEE Transactions on Power Apparatus and Systems PAS-101 (12), 4640–4648.

Chao, H.-p., Peck, S., Oren, S., Wilson, R., 2000. Flow-based transmission rights and congestion management. The Electricity Journal 13 (8), 38–58.

Chao, H.-P., Peck, S. C., 1998. Reliability management in competitive electricity markets. Jounal of Regulatory Economics 14, 189–200.

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

Dana, J. D., 1993. The organization and scope of agents: Regulating multiproduct industries. Journal of Economic Theory 59, 288–310.

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.

de Maere d'Aertrycke, G., Smeers, Y., 2013. Liquidity risks on power exchanges: a generalized nash equilibrium model. Mathematical Progamming Ser. B (140), 381–414.

Deng, S.-J., Oren, S., Meliopoulos, S., 2010. The inherent inefficiency of simultaneously feasible financial transmission rights auctions. Energy Economics 32, 779–785.

Ehrenmann, A., Smeers, Y., 2005. Inefficiencies in european congestion management proposals. Utilities policy 13 (2), 135–152.

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

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

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

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

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

European Commission, January 2014b. Impact Assessment - A 2030 framework for climate and energy policies. SWD(2014) 15 final.
URL `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). Tech. rep., EuroWind GmbH.

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.

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 (2), 125–149.

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 (3), 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.

Harvey, S. M., Hogan, W. W., January 2000. Nodal and zonal congestion management and the exercise of market power.

Höffler, F., Wambach, A., 2013. Investment coordination in network industries: The case of electricity grid and electricity. Journal of Regulatory Economics 44 (3), 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 (2), 211–242.

Hogan, W. W., April 1999. Restructuring the electricity market: Institutions for network systems.

Huppmann, D., Egerer, J., 2014. National-strategic investment in European power transmission capacity. DIW Discussion Papers, No. 1379.

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.

Joskow, P., Tirole, J., June 2005. Merchant transmission investment. The Journal of Industrial Economics LIII (2), 233–264.

Joskow, P. L., 2014. Incentive Regulation in Theory and Practice: Electric Transmission and Distribution Networks. University of Chicago Press, Ch. 5.

Kristiansen, T., 2005. Markets for financial transmission rights. Energy Studies Review 13 (1).

Kunz, F., 2013. Improving congestion management: How to facilitate the integration of renewable generation in germany. The Energy Journal 34 (4), 55–78.

Kurzidem, M. J., 2010. Analysis of flow-based market coupling in oligopolistic power markets. Ph.D. thesis, ETH Zurich.

Laffont, J.-J., Martimort, D., 2002. The theory of incentives - the principial-agent model. Princeton University Press.

Laffont, J.-J., Tirole, J., 1993. A Theory of Incentives in Procurement and Regulation. MIT Press.

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.

Lewis, T. R., Sappington, D. E., 1988a. Regulating a monopolist with unknown demand. The American Economic Review 78 (5), 986–998.

Lewis, T. R., Sappington, D. E., 1988b. Regulating a monopolist with unknown demand and cost functions. The RAND Journal of Economics 19 (3), 438–457.

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

Monitoringbericht, 2013. Bundesnetzagentur, Bundeskartellamt.

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

Myerson, R. B., 1979. Incentive compatibility and the bargaining problem. Econometrica: Journal of the Econometric Society 47 (1), 61–73.

Netzentwicklungsplan, 2013. Netzentwicklungsplan Strom 2013 - Zweiter Entwurf der Übertragungsnetzbetreiber. 50Hertz Transmission, Amprion, TenneT TSO, TransnetBW.

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 (3), 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.

Ozdemir, O., Munoz, F. D., Ho, J. L., Hobbs, B. F., 2015. Economic analysis of transmission expansion planning with price-responsive demand and quadratic losses by successive lp. IEEE Transactions on Power Systems PP (99), 1–12.

Platts, December 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.

Sappington, D., 1983. Optimal regulation of a multiproduct monopoly with unknown technological capabilities. The Bell Journal of Economics 14 (2), 453–463.

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 (123-135).

Schmalensee, R., 1989. Good regulatory regimes. The RAND Journal of Economics 20 (3), 417–436.

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

Siddiqui, A. S., Bartholomew, E. S., Marnay, C., Oren, S. S., 2005. Efficiency of the new york independent system operator market for transmission congestion contracts. Managerial Finance 31 (6).

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.

Viehmann, J., 2011. Risk premiums in the german day-ahead electricity market. Energy Policy, 286–394.