Charles University

Faculty of Social Sciences

Institute of Economic Studies

MASTER'S THESIS

Natural gas in the EU: An empirical study of price determinants in the age of blooming shale gas and LNG exports

Author: **BSc. Zankoo Abbaspour**
Supervisor: **prof. Ing. Evžen Kočenda MA, Ph.D., DSc.**
Academic Year: **2018/2019**

Declaration of Authorship

The author hereby declares that he compiled this thesis independently; using only the listed resources and literature, and the thesis has not been used to obtain a different or the same degree.

Prague, December 24, 2019

Signature

Acknowledgments

This thesis is part of a project that has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 681228.

Abstract

This paper forecasts day-ahead prices of Title Transfer Facility Gas Hub, Europe's most liquid gas market, by employing a comprehensive list of autoregressive and regionalized fundamental variables. Using a dataset containing daily data for the period 2018-2019, we estimate two specifications using Extreme Gradient Boosting Algorithm. We find that yearly differentials in filling rate of European underground gas storage to carry significant information gain. Our results also confirm short term inertia in the price.

Contents

List of Tables

List of Figures

Acronyms

LNG	*Liquefied Natural Gas*
TTF	*Title Transfer Facility (The Netherlands)*
NBP	*National Balancing Point (UK)*
NCG	*NetConnect Germany*
PSV	*Punto di Scambio Virtuale (Italian virtual gas hub)*
ICE	*Intercontinental Exchange*
MWh	*Mega Watt hour*
KWh	*Kilo Watt hour*
GWh	*Giga Watt hour*
TWh	*Tera Watt hour*
MMBtu	*Million British thermal units*
Bcm	*Billion Cubic Meter*
NWE	*Northwestern Europe*
CWE	*Central Western Europe*
CEE	*Central Eastern Europe*
SEE	*Southeastern Europe*
DA	*Day-ahead*
MA	*Month-ahead*
UGS	*Underground Gas Storage*
ALSI	*Aggregated LNG Storage Inventory*

Master's Thesis Proposal

Author:	BSc. Zankoo Abbaspour
Supervisor:	prof. Ing. Evžen Kočenda MA, Ph.D., DSc.
Defense Planned:	Dec 2020

Proposed Topic:

Natural gas in the EU: an empirical study of price determinants in the age of blooming shale gas and LNG exports

Motivation:

In 1998 European Commission voted for the Directive 98/30/EC concerning common rules for the internal natural gas market. The directive gives market participants access to 'natural gas undertakings'. These activities include import, export, transmission, storing and trading. As a physical commodity, gas is traded on different exchanges for various delivery points across the European Union (EU). The most notable of which are Title Transfer Facility (TTF) in the Netherlands, National Balancing Point (NBP) in the United Kingdom and NetConnect (NCG) in Germany. These so-called Virtual Trading Points (VTP) have different prices due to various underlying reasons. However, Growitch et al. (2012) shows high degree of interdependency and convergence among them. Prior to the market liberalization, coupling of crude oil and natural gas prices was a common knowledge. Nowadays however, the evidence points to a weakened relationship (Stern, 2014). Hub prices, nowadays more than ever, reflect the fundamentals of gas market's supply and demand. This is not the only shift in the dynamics of the market. European Union's status as a major consumer and importer of the commodity warrants a reassessment of fundamental drivers of the price in the region and introduction of new variables which in the previous market regimes were perhaps deemed irrelevant.

European Union is a globally significant consumer and importer of natural gas. As it stands it consumes 500 billion cubic meters per annum. By many estimates this figure will grow by 10% to 20% in next the five years. Even though the bulk of the growth will be dominated by the power generation end use, the residential demand remains to be the largest force. In parallel to the growing demand, however, the inland production has already peaked in 2015. Therefore, in order to reach a certain level of energy security import diversification seemed inevitable. Such diversification would lead to emergence of new price determinants. Case in point, Russian pipeline imports, which are arguably the most significant sources of import to the European markets are sold through so called oil-indexed long-term contracts of up to 25 years, are now being challenged by LNG cargos from Qatar, U.S. and other major LNG exporters. In the U.S. case, this would in theory introduce volatility of Henry Hub price index to the European markets for better or worse. Furthermore, on the back of the recent non-associated shale gas revolution in the States, the volumes entering the EU can be substantial. These ongoing trends have led to a new paradigm in the European gas market. A market where its price determinants need to be re-evaluated irrespective of crude oil prices.

In their findings, Hartely et al. (2008) show traditional cointegration of gas prices with crude oil has become less significant. There are many ongoing shifts reshaping the dynamics of the European natural gas market such as growth in renewable energy generation, expanding LNG import and export infrastructures, pipeline import diversification, diminishing of the so-called Asian gas premium etc. to name a few. Each of the mentioned trends can contribute to the explanation of the price volatility in today's natural gas market. In the light of such developments,

it is of interest to reassess the explanatory power of classic price drivers and also to introduce and evaluate new variables that have recently gained relevance.

Hypotheses:

Hypothesis #1: Gas market fundamentals contain sufficient information gain to explain short-term price action without resorting to crude oil prices
Hypothesis #2: Renewable energy generation does not have a significant impact on the European natural gas spot prices

Methodology:

In the first phase, a list of potential fundamental variables will be introduced and discussed. Nick and Theones (2016) included temperature, supply shortages and storage capacity in their short-term model. Whereas in Vector Error Correction Models, typically geared towards longer term fluctuation modelling, coal and crude oil prices are almost universally present. Later, some other variables that are theoretically deemed capable of shifting the supply demand equilibrium will be added to that list. These price determinants have been either discussed by the previous works in relation to other markets than EU or not mentioned at all.

The problem tackled in this paper revolves around the idea of estimating the day-ahead gas prices using gradient boosting algorithm based on fundamental and autoregressive variables. Fundamental variables are believed to play a quintessential role as they affect the balance of the gas market. In this context, balance of the market refers to the difference between the expected demand volumes and the available supply capacity for the day-ahead delivery. On the other hand, the autoregressive variables are expected to similarly account for a major part of price variation mainly due to momentum driven nature of gas prices that has partially to do with inertia associated with weather.

This paper employs the efficient 'Extreme Gradient Boosting' Algorithm for the first time to forecast European gas spot prices on an up-to-date data set spanning from Dec. 2018 to Oct. 2019. This in itself a major contribution as it re-evaluates the significance of relationships discovered by previous studies in almost a decade ago in the new environment of booming LNG and shale gas production. This paper also accounts for all the significant production of solar and wind energy, gas and LNG storages, and pipeline flows throughout Europe and introduce them into the model in regionalized form. Such level of comprehensiveness for non-linear regression analysis in the literature is unprecedented.

Expected Contribution:

- ! Use of Extreme Gradient Boosting Algorithm for the first time to estimate short-term European gas prices on an extensive list of fundamental and autoregressive inputs
- ! Using a dataset that is both current and geographically comprehensive

Outline:

1. Motivation: I will explain the framework of Natural gas market in the EU and on a global scale. Moreover, I will point out the current trends in the supply chain and demand prospect with focus on the EU.
2. Literature review: I will introduce a set of fundamental variables discussed by the core literature.
3. Data: I will explain how and what I will collect. Also, I will discuss the feature engineering techniques performed on selective variables.
4. Methods: I will explain the estimation techniques and performance metrics.
5. Results: I will discuss and compare my models' results and benchmark their performance against similar works in the field.
6. Concluding remarks: I will summarize my findings and conclude if the hypotheses have been rejected

Core Bibliography:

Busse, S., Helmholz, P., & Weinmann, M. (2012). Forecasting day ahead spot price movements of natural gas – An analysis of potential influence factors on basis of a NARX neural network. *Braunschweig: Institut für Wirtschaftsinformatik*.

Čeperić, E., Žiković, S., & Čeperić, V. (2017, 12). Short-term forecasting of natural gas prices using machine learning and feature selection algorithms. *Energy, 140*, 893-900.

Crow, D. J., Giarola, S., & Hawkes, A. D. (2018, 5). A dynamic model of global natural gas supply. *Applied Energy, 218*, 452-469.

Daan Hulshof, M. (2016, 7). Market fundamentals, competition and natural-gas prices. *Energy Policy, 94*, Pages 480-491.

Emma Schultz, J. S. (2013, 10). Price discovery in European natural gas markets. *Energy Policy*.

Geng, J.-B., Ji, Q., & Fan, Y. (2016). The impact of the North American shale gas revolution on regional natural gas markets: Evidence from the regime-switching model. *Energy Policy, 96*, 167–178.

Growitsch, C., Stronzik, M., & Nepal, R. (2012). Price Convergence and InformationEfficiency in German Natural Gas Markets. *German Economic Review*.

Hang T. Nguyen, I. T. (2010). Short-term electricity demand and gas price forecasts using wavelet transforms and adaptive models. *Energy, 35*, 3674-3685.

Hartley, P. R., Medlock, K. B., & Rosthal, J. E. (2008). The relationship of natural gas to oil prices. *Energy Journal*.

Heather, P. (2010). The Evolution and Functioning of the Traded Gas Market in Britain.

Junghwan Jin, J. K. (2015, 11). Forecasting Natural Gas Prices Using Wavelets, Time Series, and Artificial Neural Networks. *PLoS ONE*.

Nazim Regnard, J.-M. Z. (2014). What drives natural gas prices? — A structural VAR approach. *Energy Economics*.

Ramberg, D., & Parsons, J. (2012). The weak tie between natural gas and oil prices. *Energy Journal, 33*, 13–35.

Sebastian Nick, S. T. (2014). What drives natural gas prices? — A structural VAR approach. *Energy Economics*.

Stern, J. (2014). International gas pricing in Europe and Asia: a crisis of fundamentals. *Energy Policy*.

Stopa, J., & Kosowski, P. (2018, 7). Underground Gas Storage In Europe-energy Safety And Its Cost.

Vedat Yorucu, P. B. (2015, 5). Price modelling of natural gas for the EU-12 countries: Evidence from panel cointegration. *Journal of Natural Gas Science and Engineering, 24*, 464-472.

Wang, T., Zhang, D., & Broadstock, D. C. (2019). Financialization, fundamentals, and the time-varying determinants of US natural gas prices. *Energy Economics*.

Zhang, D., & Ji, Q. (2018). Further evidence on the debate of oil-gas price decoupling: A long memory approach. *Energy Policy*.

Author **Supervisor**

1. Introduction

Against the backdrop of a decrease in investments in the mid 80's, the European Commission reacted with a series of directives aiming at the liberalization of the energy sector to ensure firstly maximized social welfare through market-driven energy prices and secondly supply diversification. Even though, some argue the second objective has failed to materialize, yet the process of liberalization is in full swing spreading across 22 members, as of writing of this paper, with Portugal being the latest country to have liberalized its market in 2011. In 1998, European Commission voted for the first energy package enshrined in Directive 98/30/EC (EC, 1998) formulating common rules for the internal natural gas market. It set out the initial steps towards changing the industry structure and network access conditions by introducing legal unbundling and negotiated and regulated third party access. Subsequently, through a process called vertical unbundling independently run gas transmission operators were created who had no vested interest in the upstream sector. The directive also gave market participants access to 'natural gas undertakings.' These activities include import, export, transmission, storing and trading. At the core of this initiative was the implementation of the "Entry-Exit-Model" unveiled in 2005. This model was based on the creation of large market areas, which can be thought of as gas pools, in which one has to agree on one entry and one exit point. By implementing this system, the barriers for supplying gas to customers nationwide were lowered significantly, promoting competition. However, it was only after the third energy package, Directive 2009/73/EC (EC, 2009), that the common rules for a liberalized market were put in full effect and enforced. The main emphasis of the latest directive was aimed at catering for indiscriminate network access to all third parties, consumers and producers alike. In doing so, the prices would reflect market fundamentals and were safeguarded from supplier manipulation. Despite vertical unbundling, the transmission networks did not have a clean start as they inherited long term legacy contracts that needed to run their course before making those booked capacities available to the market. Most of these contracted capacities went hand in hand with Take-or-Pay gas contracts whose price was linked to oil or some basket of petroleum products.

In 2005, nearly *80%* of the gas sold in Europe was priced in relation to oil vis a vis only *15%* share of hub pricing. By 2013, gas priced in relation to oil had fallen to just over *40%* and more than *50%* of the volumes was market priced. These figures were

even more dramatic in Northwestern Europe with gas-on-gas share at *80%* and trending higher (Stern & Rogers, 2014). In other words, the price discovery in Northwestern Europe is increasingly a function of the gas market fundamentals rather than oil's nor is it an outcome of game theories dictated by handful of major suppliers. To gain insight into European hub price dynamics, one must, in the first place, identify sources of supply with similar characteristic, i.e. pricing and flow dynamics, to be grouped according to flexibility; an important price driver. One way of doing this is the bottom-up approach, where every field's long-run marginal cost is accounted for. This approach is characteristic of dynamic linear programming models and requires tremendous amount of resources. Moreover, the validity of such approach is called into question as more and more of North American and Asian LNG cargoes are pouring into the continent. Alternatively, given the few numbers of suppliers, i.e. Russia, Norway, Libya, and Algeria, to Europe, one can simplify the problem by geographically aggregating pipeline flows, a method exercised by this paper.

As a physical commodity, gas is traded on different exchanges for various delivery points across the European Union. The most notable of which are Title Transfer Facility (TTF) in the Netherlands, National Balancing Point (NBP) in the United Kingdom and NetConnect (NCG) in Germany. NBP and TTF are, in turn, the most liquid European hubs for both short-term, i.e. day-ahead delivery, and longer-term futures products in Over the Counter and Exchange Cleared Markets. A day-ahead product is a physical contract of certain quantity, in MWh, that is traded one day before the delivery date and delivered from 6 a.m. of the delivery day until 6 a.m. of following day. For instance, a 1 MWh day-ahead contract, delivers 1MW per each hour for a 24-hour duration.

The so-called Virtual Trading Points have different prices due to various technical underlying reasons. However, Growitch et al. (2012) show high degree of interdependency and convergence among them. Prior to the market liberalization, coupling of crude oil and natural gas prices was a common knowledge. Nowadays however, the evidence points to a weakened relationship (Stern, 2014). Hub prices, nowadays more than ever, reflect the fundamentals of gas market's supply and demand. This is not the only shift in the dynamics of the market. Europe's supply diversification has been accelerating in recent years towards adopting more LNG and non-Russian pipeline volumes. Such strategic shift warrants a reassessment of fundamental drivers of the price in the region and introduction of new variables which in the previous market regimes were perhaps deemed irrelevant. Nowadays, with diminishing share of oil-linked contracts, more and more of the new long-term

contracts include at least one hub price in the price formula. Therefore, it stands to reason that the share of hub pricing should and have increased with time. The forecast of future development of the natural gas prices, therefore, has gained more relevance.

Now that a decisive share of gas volumes is sourced in short term markets benchmarked against hub prices, it pays to be able to forecast the fair value of the commodity. The task has become more tractable since these hub prices reflect the fundamentals of the market and especially so in the most liquid hubs i.e. TTF. Accurate forecast enables market participants to manage their portfolio optimally by placing bids and asks at the right price levels helping them to outperform their opponents and reach optimal portfolio gain. The first step in achieving a reliable forecast is knowledge of the price determinants. The problem tackled in this paper revolves around the idea of estimating the TTF day-ahead gas prices using Extreme Gradient Boosting Algorithm based on fundamental and autoregressive variables. Fundamental variables such as physical pipeline flows transiting gas into the continental Europe, LNG and Underground Gas Storages, renewable energy production etc. are believed to play a quintessential role as they affect the balance of the gas market. In this context, balance of the market refers to the difference between the expected demand volumes and the available supply capacity for the day-ahead delivery. On the other hand, the autoregressive variables are expected to similarly account for a major part of price variation mainly due to momentum driven nature of gas prices that has partially to do with inertia associated with the weather.

The data used in this paper is taken from various sources. A part of it is sourced directly from original sources such as European Network of Transmission System Operators for Gas, Gas Infrastructure Europe, Eurostat, etc. The rest of the data is sourced from second hand data vendor such as Refinitiv and Wattsight. Data processing mostly entails aggregation, unit standardization and filtering. The aggregation process is two-faceted: geographical and temporal. A novel contribution by this paper is the use of Extreme Gradient Boosting Algorithm for estimation purpose. In the estimation phase, the dataset is divided into train and test splits, whereby, at each new fold a test data set is forecasted using a boosted estimator, a decision tree, parametrized on the preceding train data set. Extreme Gradient Boosting Algorithm, known as ‘XGBoost’, is a supervised learning method that is based on function approximation by optimizing a specific loss function as well as applying regularization techniques. It is an enhanced implementation of gradient boosting that is both computationally efficient in dealing with high-dimensional datasets and immune to the overfitting problem.

This paper employs the efficient XGBoost algorithm for the first time to forecast European gas spot prices on an up-to-date dataset spanning from Dec. 2017 to Oct. 2019. The scope of data covered in the literature is out of date. In this regard, this paper brings a major contribution into the literature as it re-evaluates the significance of relationships discovered by previous studies in almost a decade ago in the new environment of booming LNG and shale gas production. This paper also accounts for all the significant production of solar and wind energy, gas and LNG storages, and import pipeline flows into the Europe and introduces them into the model in regionalized form. Such level of comprehensiveness for non-linear regression analysis in the literature is unprecedented.

This paper is structured as follows: first we compare and contrast a portion of the literature in gas price modelling field. In doing so, we review a handful of papers from various angles and point out to the corresponding strengths and short comings, upon which this paper improves. In the later section, we introduce the theory behind the XGBoost algorithm and its cutting-edge features. In the same section, we also introduce the concept of information gain, as a tool to identify important features, and performance metrics used later at the performance evaluation part. In the next section, we describe the factors that contribute to the change in the balance of the market from both fronts: demand and supply. Upon listing each factor, we proceed to explain and justify their correlation and relevancy. Having introduced and explained the rationale behind each input, we then spend some time going over the dataset used in the estimation. We introduce all the input variables by name and explain the corresponding feature engineering involved in creation of each of them. Data description section is followed by results section where the findings are presented. Different model specifications and associated performances are illustrated. Later the most important features of the best performing model are shown according to their contribution to the total information gain. In the next step, the results of OLS regression of the most important set of inputs against the target variable is presented in a table. The discussion of the economic sanity of the most important feature comes next. Finally, this paper finishes with the conclusion section where findings and contributions are listed again curtly.

2. Literature review

The purpose of this section is to compare and contrast an important portion of the literature in gas price modelling field. In doing so, we review a handful of high-ranking papers from various angles and point out to the corresponding strengths and short comings, upon which this paper intends to improve. At first, we will have a look at the scope of data and examine the temporal granularity and range. Secondly, we will have a look at the number and category of explanatory variables as well as the estimated variables. And lastly, we will review methodology and the significance of their findings.

Prior to diving into examining each individual source on their merits, it is important to paint a picture describing the underlying trends in gas pricing field and its evolution since early 2000 up to date. Roughly speaking, most of the literature before and around 2009, which coincides with the maturing of Shale gas revolution, engages with the topic through conducting cointegration analysis between regional gas prices and regional oil prices. The most common methodology for this type of analysis was Vector Error Correction Model and its enhanced versions. This was mainly due to the fact that up until that period most of the upstream gas projects were mostly financed through long term purchase agreements, up to 25 years tenure, the price discovery mechanism for which was indexed to oil (EC, 1998, p. 19). In the same period, the widespread of heating-oil burning devices at the retail end made the substitution with gas practical and justifiable (Stern & Rogers, 2014). Therefore, the rationale behind the hypothesis of such studies made sense. However, after and around 2009, when the production of non-associated non-conventional gas surged, even though this was only true of the United States of America, we can observe a subtle change of course from cointegration analysis towards modelling gas prices based on own fundamental factors namely demand and supply. As the European gas hubs entered a mature state and numerous long-term oil-indexed contracts were approaching the end of their lifetime, more and more new contracts were indexed to local hubs (Yorucu & Bahramian, 2015) which reflected the gas market fundamentals more accurately than oil price did. So, it was only natural to study gas price development through its own supply and demand reality. All in all, we can summarize the gas pricing efforts into two eras, one that tries to reconcile gas prices with oil prices and one that wants to attribute it to market fundamentals. In the coming sections we focus on each individual paper and its contribution to the field and contrast them to the innovations

our paper is set to bring forth.

The scope of data covered in all the sources reviewed is out of date. Majority of the sources use daily temporal granularity except for Schultz & Swieringa (2013), who use 10-minute frequency data for the period 2008-2011, and Jin & Kim (2015), who use weekly data for the period 2000-2013. The most recent study was found to be that of Geng et al. (2016). They conducted a cointegration study on both European and North American gas indices, in turn NBP and Henry Hub, for the period 1998-2015. Hulshof et al. (2016) however carried out the most recent regression analysis of the TTF gas prices based on exogenous variables. It is interesting to note that in the 2000s, Zeebrugge and NBP gas price indices had been used as the representative hub prices for Europe. However as shown by Heather (2012), TTF, as of recently, has multiple fold more depth, churn rate and liquidity than NBP and even more so than Zeebrugge hubs in short term futures market; hence a better representative of market fundamentals for Europe. In our study, we have chosen TTF spot prices as the estimated variable. Another contribution by this paper is the use of current data. Our dataset spans from 2017-2019. This in itself a major contribution as we can re-evaluate the significance of relationships discovered by previous studies in almost a decade ago on a different hub.

In terms of explanatory variables, three aspects will be discussed: types, engineering and selection procedure. Let us begin by comparing the types of explanatory variables across the sources. In cases where Vector Error Correction Model is applied, most typically the oil price is the only explanatory variable. For instance Geng et al. (2016), use West Texas Intermediate Crude Oil Index for Henry Hub gas price and Brent Crude Index for the National Balancing Point gas price estimation. Among all the studies whose methodology was based on studying cointegration between gas and oil, only Regnard & Zakoïan (2014) introduced an exogenous variable namely temperature into the model specification.

Among linear and non-linear regression analyses, there are significantly more exogenous variables used. Regarding papers which employed some sort of regression tool to model gas prices, an abundance of exogenous variables was employed. Not only the contemporaneous but also the lags of those variable are often present in the model specification. A variable that makes frequent appearances across different studies is temperature. Very rarely temperature is entered into the model definition in its absolute raw value except in Nguyen & Nabney (2010). In both Čeperić et al., (2017) and Hulshof et al. (2016), for example, Heating Degree Days and Cooling Degree Days have been constructed to proxy for temperature. These measures are

essentially differentials between observed outside temperature and a jurisdiction-specific threshold below or above which the central heating or cooling for buildings would be triggered. In these studies, temperature is believed to be a proxy for gas consumption. This is of course a solid line of thinking; as in Europe as well as in North America the bulk of gas consumption goes to heating energy generation. Nguyen & Nabney (2010), however, include both temperature and gas demand in United Kingdom's gas price estimation. One might object to the representation of consumption solely through temperature, as not all the consumption stems from heating need but also from industrial and commercial activities. Based on that point, weekends have different consumption profile and temperature alone is not a wholesome representative proxy. Since the macro data regarding industrial activity is published in monthly granularity, to remedy the problem of distinguishing between consumption behavior on work days and weekends Hulshof et al. (2016) and Nguyen & Nabney (2010) use time dummies to account for different days of the week. Similarly, Čeperić et al. (2017) use time dummies to distinguish both days of the week and months of the year. In this paper, we address the seasonal variation of consumption by using weekly and yearly differentials of all exogenous variables.

Aside from temperature, renewable energy production can also influence the consumption pattern. With the rise of renewable energy production and rising coal switching behavior (Timera Energy, 2016), characteristic of current low gas price and high emission prices in Europe, wind and solar energy production can in theory affect gas demand. Having said that, most studies ignored the potential correlation or only partially accounted for it. In their estimation of TTF gas prices, Hulshof et al. (2016) uses German wind generation as a proxy for renewable energy generation in North Western Europe. Even though, geographically the weather across this region shows high level of homogeneity, yet exclusion of the wind and solar production from other states that rely on TTF hub for their gas supply is a point of concern. Nguyen & Nabney (2010), on the other hand, goes one step back and only includes average UK wind speed in their model. Our approach is more thorough and comprehensive. More specifically, we will account for all significant production of solar and wind energy throughout Europe and introduce them into our model in regionalized aggregated form.

In addition to contemporaneous exogenous variables, their associated lags and autoregressive components of the estimated variable can be also found in the mix of explanatory variables. Schultz & Swieringa (2013) use month-ahead prices of various futures products from various European exchanges and their lags of up to 10 to estimate Zeebrugge spot prices. In attempt to gain more sparsity, they filtered out

irrelevant lags of high correlation variables using Automatic Relevance Determination Regression. In this method the variables with near zero Ridge coefficients were removed from the model. Čeperić et al. (2017) take the process of adding relevant lags one step further by automating it. Using a feature selection technique, known as 'Steepwise Feature Selection', new inputs are generated by introducing lagged values of the exogenous variables in sequential stepwise regression. In other papers where the method of estimation was based Wavelet decomposition i.e. Jin & Kim (2015) and Nguyen & Nabney (2010), detail components of the estimated variables, in turn UK monthly gas forward contracts and Henry Hub spot prices, were added as explanatory variables. As mentioned before, in this paper, we will add weekly and yearly differentials of all exogenous variables into mix. Moreover, autoregressive components TTF prices up to lag order of 7 will be present in the list of inputs. We will evaluate importance of each feature based on the concept of 'Percentage Gain', which will be discussed in the theory section. Since the estimation algorithm of choice in this paper is highly scalable, all the inputs will be fed to the model without pre-selection and feature importance analysis is only carried out ex-post.

Two general approaches to gas price forecasting are regression and linear programming methods. The sources discussed so far do entirely belong to regression analysis of different variety such as linear, non-linear, error-correction, etc. On the other hand, linear programming studies are less diversified and clustered around a well-established model usually developed by a well-known establishment.

In Europe, for instance, TIGER natural gas infrastructure and dispatch model or its variations are employed in numerous studies simulating regional flows and prices. This model was developed by Lochner & Bothe (2007) at the Institute of Energy Economics, University of Cologne. Most current studies, which take the linear programming path, apply TIGER model or its enhanced versions. The reason for such seemingly lack of creativity is mostly due to resource intensity of building a similar model. TIGER or models of this sort belong to fundamental category where essentially the gas infrastructure, up to a certain degree of details, including transmission grid, consumption areas, supply sources are modeled into a nodal network. By applying certain demand and supply elasticity assumptions and then subjecting them to grid bottlenecks, an optimal solution i.e. price and flow for each node and interconnection is found while maximizing social welfare. An exact replica of the actual real-life infrastructure should in essence provide a one-to-one simulation of the market supposing the market is efficient. A highly detailed fundamental bottom-up models is immensely resource intensive. Therefore, the goal is to strike a

balance between having a simplified yet accurate representation of the real-life infrastructure. This can be done by regionalizing demand and supply areas. Since high resolution models need to both take into account large amount of data and maintain the currency of assumptions for each individual infrastructural unit, they are usually sponsored through longer-term projects on institutional level. Of course, once one is built, it makes more sense to improve on the existing model rather than building parallels to that; since, there is little room for subjective interpretation of, for example, capacity of pipeline connecting Germany to the Netherlands. Lochner (2011) uses an enhanced version of TIGER to simulate the European natural gas market during the 2009 Russian-Ukrainian gas conflict. Simulation results showed high resemblance to the actual physical flow of gas post-crisis. This means the industry reacted with a high level of efficiency. The largest contribution to compensating the Ukrainian missing volumes came from storages, which was the case in the reality as well. This highlighted the critical role of gas stocks in the face of supply shortages. We account for all the underground gas and LNG storage capacity across Europe in this paper. Similarly, we also report that storages play the most significant role in estimating the short-term prices.

Regression methods, on the other hand, are more affordable to experiment with. Among studies opting to analyze gas prices using such methods, one observe an extensive array of methods and creativity. Regnard & Zakoïan (2014), for instance, found out single regime is too restrictive. Temperature varying Autocorrelation Function showed GARCH model's fixed coefficients were too rigid as the behavior varies according to the temperature regimes. Instead, they opted for three different temperature brackets , , and . They also fitted a 4-regime for each quarter of the gas year. They found out the volatitly of second temperature regime is less persistent than the other two. In that particular regime, the impact of recent obseravations on the volatility is stronger than the low and high temperature regimes, which indicates headline driven nature of the market. regimes showed more resemblance as did regimes. Winter season also showed high degree of inertia in respect to recent price developments. These results show the importance of seasonality in the gas prices. Ordinary Least Square technique would fail to recognize regime changes; however, a decision tree is more than capable of detecting those key thresholds associated with behavior changes.

In their high frequency study of Zeebrugge price determinants using Multivariate Vector Error Correction Model, Schultz & Swieringa (2013) reported that lagged returns of all month-ahead NBP, day-ahead TTF and ICE monthly contracts were significant. Also, low value coefficients of contemporaneous variable suggested

market friction among various hubs. Contemporaneous coefficient of day-ahead products were higher. As they also pointed out, this has to the overarching nature of heating demand over a closely connected geographical area namely Northwest Europe. Their results also reported, NBP month-ahead and ICE monthly futures have strong explanatory power on short term development of other securities. They concluded most important contribution to the innovation of price in longer term equilibrium comes from NBP and then TTF day-ahead prices. This and many other studies have already proven the large degree integration across major European hubs. There seems to remain some friction among the hubs on the back of infrastructure bottlenecks. In our model, we will add both contemporaneous and lagged timeseries of a hand full of major hubs including NBP and PSV to the list of explanatory variables.

In two studies where non-linear methods are the main forecasting tools, the results pointed to performance improvement when wavelet components are present in the model specification. Nguyen & Nabney (2010) transformed target variable, daily price of gas forward contracts, using redundant wavelet transformation. Each wavelet component along with exogenous variables was then forecasted separately using either of Multilayer Perceptron (MLP), Radial Basis Function (RBF), Linear Regression and GARCH methods. Linear regression and GARCH outperformed non-linear MLP and RBF models. Multicomponent forecast outperformed models without wavelet transformation. Also, Adaptive models delivered higher performance. For instance, adaptive GARCH models with wavelet components achieved the best results. Similarly, Jin & Kim (2015) found out that wavelet decomposition combined with ARIMA estimation was the best choice for longer term forecasting of Henry hub spot prices. They forecasted the prices using wavelet, timeseries models and Artificial Neural Networks (ANN). In their paper, the most influential factor in obtaining higher performance in regard to ANN was the number of hidden nodes. Two nodes delivered most accurate results without converging and three nodes converged with less than desirable results. In gradient boosting setting, these set of parameters corresponds to number of terminal nodes, whereby the process of fine-tuning them is known as ‘hypertuning’. They also found out that the inclusion of wavelet Detail component only slightly improved the results. Whereas, wavelet decomposition combined with ARIMA was the best choice for longer term forecasting. Their results hinted at better suitability of GARCH models for forecasting the detail components. Decision tree are rather robust to noisy data. Put in other words, threshold values do not distinguish between a true value or its distorted version so long as information gain is maximized. Therefore, it is safe to assume denoising data will not bring significant performance enhancement. Hence, in our analysis, we will not use any

decomposition technique and the price time-series will be modeled in its raw form. In doing so, the processing will be more straightforward and computationally more affordable.

3. Theory

3.1. Gradient boosting and function estimation

For a given set of data with as the output variable and as the input set of variables, a portion of the data , also called the training sample, is used to estimate a 'mapping' function that relate to y. The function is approximated as such that a desired loss function for the training sample is minimized. The minimizing mapping function is defined as follows:

One of the most common loss functions is squared-error .

One way of arriving at the optimal estimating function is through additive expansion of a series of parametrized sub-optimal estimators , where refers to group of parameters defining each function, also known as learners in the machine learning jargon, defined as:

Function refers to a very basic parametrized learner by a set of parameters . As each individual function gets parametrized through an iterative random sampling of the dataset, it is more likely that the set of parameters differ from one learner to another. In this paper, the candidate learner is a regression tree that is boosted sequentially through a method called Extreme Gradient Boosting. In the regression tree case, the set of parameters comprises of number of leaves, split variables and associated thresholds.

In essence the problem of estimating the parametrized function boils down to estimating the optimal parameters. More specifically, the parameters are optimized through incremental additions to our initial guess as such:

In the steepest-descent method, in order to compute each addition, , we need to calculate gradient as follows:

where

The incremental boost then is computed as such

Where

According the Friedman et al. (1983), we define a J-terminal node regression tree as follows:

where represents all the terminal nodes that accommodate all range of the explained variable. The indicator function I(.) take binary values of one or zero depending on whether or not belongs to terminal node . represents the set of parameters i.e. splitting variables and associated thresholds, defining boundaries of . The gradient-boosted version of a regression tree looks like the following:

3.2. XGboost Algorithm

XGBoost is an implementation of gradient boosting. In the following paragraphs, we go through the two of its most important innovations that have contributed to its recent popularity. Firstly, XGBoost has an option to penalize complex models through both and regularization. Regularization helps preventing overfitting problem. Missing values or data processing steps like one-hot encoding makes the data sparse. XGBoost incorporates a sparsity-aware split finding algorithm to handle different types of sparsity patterns in the data.

We want the model to be as accurate as possible. It's also important to add a

regularization term to prevent overfitting, a situation where the model ends up memorizing the sample dataset, and as a result, does not perform accurately on a new sample. The objective function (loss function and regularization) at iteration that we need to minimize is the following:

(1)

where

It is easy to see that the objective is a function of functions, i.e. is a function of Classification and Regression Tree learners, and as Chen & Guestrin (2016) point out it "cannot be optimized using traditional optimization methods in Euclidean space." The loss function in XGBoost Algorithm has a regularization term penalizing the over-complexity of the tree structure. Whereby, refers to the number of leaves in each tree and refers to the corresponding weights. This loss function is used to calculate maximum gain which can be directly used in tree building process while splitting each node. So, this is how the two-step process in gradient boosting, i.e. 1) calculation of pseudo-residuals and 2) calculating incremental value , is reduced to one single step with regularization. Moreover, the algorithm can automatically learn how to best handle missing data. In fact, it was designed to work with sparse data i.e. one-hot encoded and missing data are handled the same way that sparse or zero values are handled; by minimizing the loss function.

When building a decision tree, a challenge is to decide how to split a current leaf. A 'greedy' (Friedman et al., 1983) way to do this is to consider every possible split on the remaining features and calculate the new loss for each split; you could then pick the tree which reduces the loss the most. By using second order approximation we can quickly optimize the objective function in (1), where at each split the objective functions simplifies to:

(2)

Where

and are the first and second order gradients of the loss function. Subsequently, we can expand the regularization term in (2) in the following manner:

(3)

(4)

Where we define as a set encompassing all the values inside leaf . Also, refers to the decision function that connect instance of to a certain leaf and . From (4) an instance of optimal leaf weight can be calculated as follows:

And finally, in the last step, we arrive at a scoring function that can help us benchmark trees of different structure and choose the optimal split variables and associated splitting threshold as follows:

. (5)

Of course, in practice calculating the score of all possible parameter combination is extremely expensive. Therefore, greedy algorithm is employed to expand the tree from a single leaf iteratively onwards. More specifically, each split yield a set of left and right , , instances of the explained variable. For that particular split, the score is calculated as follows:

] - (6)

3.3. Feature importance

In fitting the XGBoost model an Importance Matrix will be produced. The importance matrix is a table with the first column including the names of all the features actually used in the boosted trees, the other columns of the matrix are the resulting 'importance' values calculated with different importance metrics. One of the importance metrics in that table is 'Gain', which implies the relative contribution of the corresponding feature to the model calculated by taking the feature's contribution for each tree in the model. A higher value of this metric when compared to another feature implies it is more important for generating a prediction. In this paper, we rank

the input variable importance using the percentage of total contributed gain across the whole ensemble of trees.

3.4. Performance evaluation metrics

In this paper we employ three metrics to evaluate the performance of the two model configurations: Mean Absolute Error, Mean Squared Error, and R-squared. Below are the description and definition of each metric.

Mean Absolute Error and Mean Square Error are the two of the most common metrics applied to measure accuracy for continuous variables. Mean Absolute Error measures the average magnitude of the errors in a set of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

Mean Square Error is a quadratic scoring rule that also measures the average magnitude of the error. It's the average of squared differences between prediction and actual observation.

Both MAE and MSE express average model prediction error in units of the variable of interest. Both metrics can range from zero to infinity and are indifferent to the direction of errors. They are negatively-oriented scores, which means lower values are better. Taking the average of squared errors has some interesting implications for MSE. Since the errors are squared before they are averaged, the MSE gives a relatively high weight to large errors. This means the MSE should be more useful when large errors are particularly undesirable. In this paper these scores are expressed in negative values; the interpretation of their absolute values, therefore, follows the same logic.

R-squared, also known as the coefficient determination, defines the degree to which the variance in the dependent variable (or target) can be explained by the independent variable (features). Put in other words, R-squared judge the goodness of fit. In this paper we have defined it as follows, where it can have negative values:

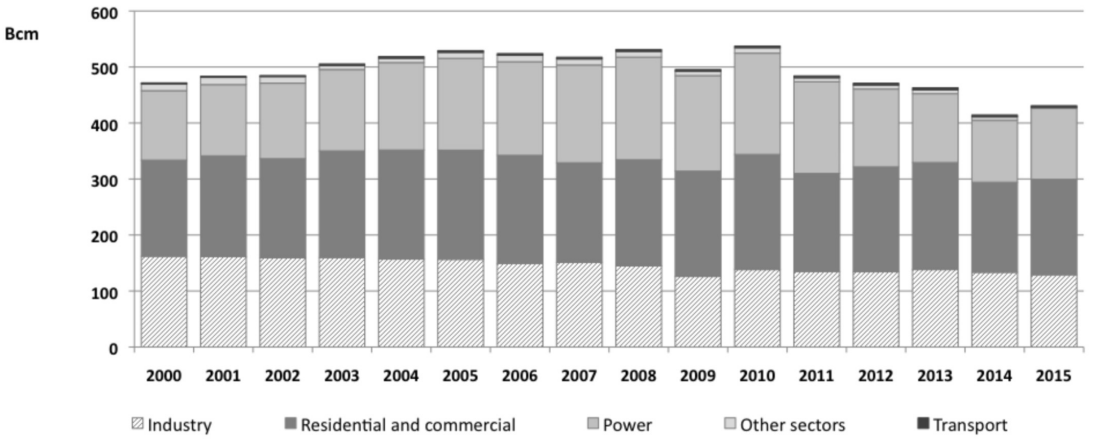
4. Balance of the market

The methodology in this paper revolves around the idea of estimating the day-ahead gas prices using gradient boosting algorithm based on fundamental variables that affect the balance of the gas market. In this context, balance of the market refers to the difference between the expected demand volumes and the available supply capacity at disposal for the day-ahead delivery. An imbalance whether caused by changes in demand or supply leads to a proportional shift in the fair or equilibrium market value of the commodity. Mainly because of geological reasons, gas supply resources around the world have different development cost and hence vary in their long run marginal cost. Qatar's North Field, for example, has a break-even point of *1.6* USD/MMbtu approximately a third of what it would cost for Norway to produce the same amount (Fattouh et al., 2015). In line with the law of supply, in the gas market for each positive incremental change in demand the market balances itself through bringing online the next competitive supply source which costs either as much as the previously balancing source or more and vice versa if demand suddenly falters. Likewise, outage of a cheap source needs be substituted by a more expensive one as in an efficient market all the cheaper alternatives must be already running at full capacity. Therefore, through these mechanisms and their reverse cases equilibrium prices can shift in both directions.

In the next sections we try to identify the factors that contribute to the changes in the balance of the market from both fronts: demand and supply. Upon listing each factor, we proceed to explain and justify its correlation and relevancy and introduce a variable or set of variables along with corresponding calculation methodology that proxy for that factor. In the preprocessing stage, once the fundamental factors are shortlisted and accounted for, their associated contemporaneous and differential variables along with lags of the endogenous variable will be used in the estimation phase.

5. Demand

5.1. Heating demand

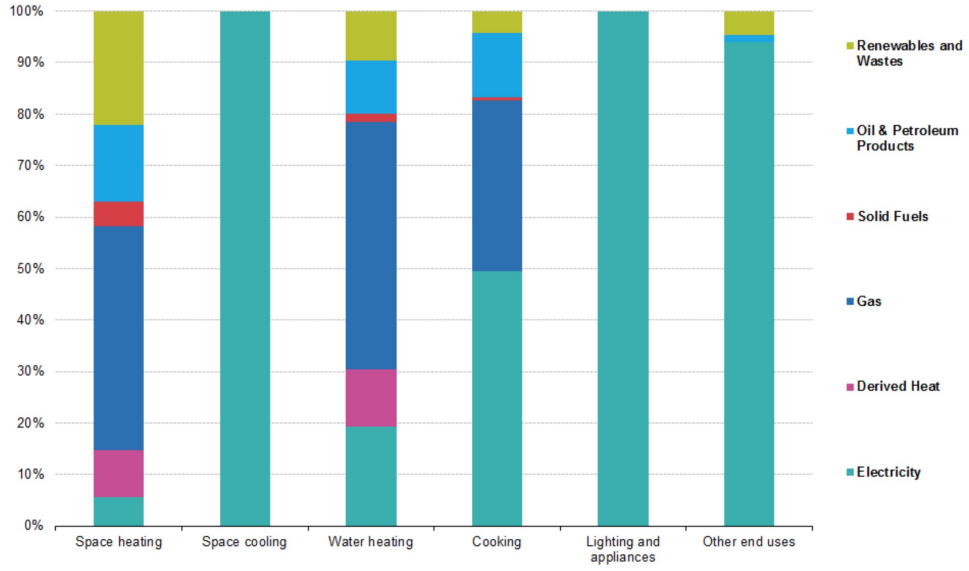


As of 2016, households accounted for nearly half of the natural gas consumption in EU-28, a trend dating back to the early 2000s (see Error: Reference source not found). In turn, space heating, water heating, and cooking are the primary end-uses by the residential sector. For space and water heating purposes natural gas is the dominant energy product; whereas for cooking, electricity is used to a wider extend (see Error: Reference source not found). Moreover, unlike the need for cooking, the amount of heating energy demanded for the other two components is directly impacted by the outside temperature and therefore not constant. As such, a temperature-based indicator can be used to model the energy consumption for these purposes. Heating degree days (*HHD*) is one of such measures.

Source: (Pisca, 2017)

HDD is defined in relation to a baseline outside temperature, below which a space i.e. a residential building requires heating. In their methodology, European Environment Agency uses *15.5 ! C* as the baseline temperature (EEA, 2016). HDD of a single day is expressed as the absolute difference between daily mean or minimum and the baseline temperature. In the monthly case, the average difference is multiplied by the number of days of in the month. The higher the value, the greater would be the demand for heating. There is a high degree of correlation between HDD and natural gas consumption in Europe (see Error: Reference source not found). In order to more

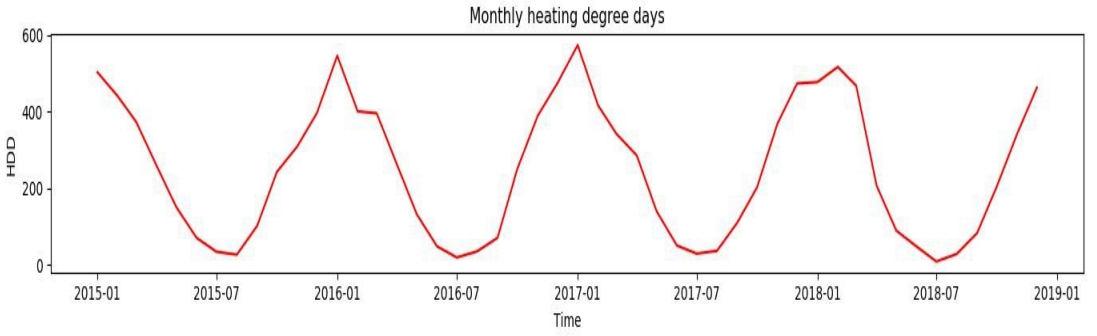
accurately capture the heating effect on the consumption, the paper by Čeperić et al. (2017) employed a three state dummy variable where temperatures below *16.5 ! C* corresponded to regular heating and the ones below *5 ! C* corresponded to extra heating effects. In their northwest European short term price forecasting model Hulshof et al. (2016) proxied consumption using a

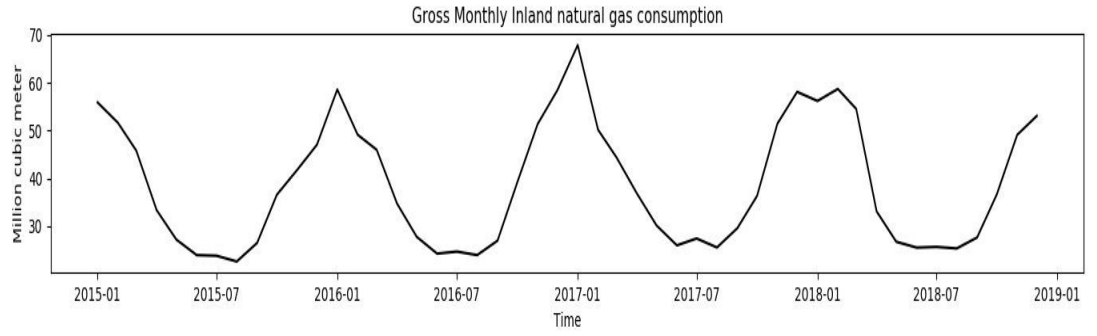


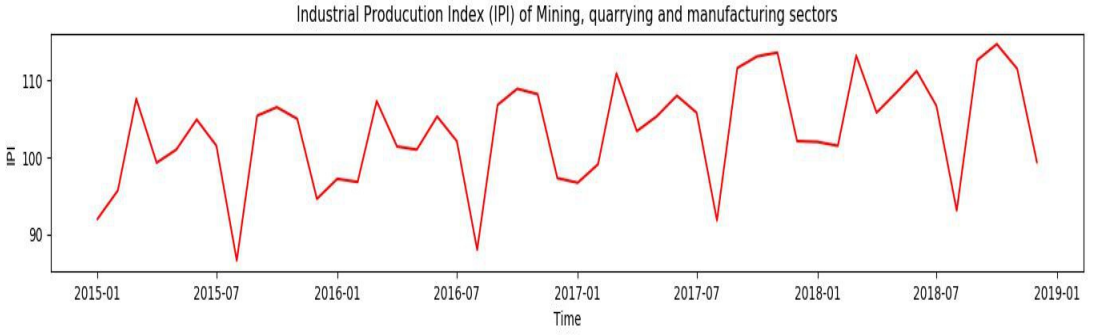
populati

on weighted average HDD.

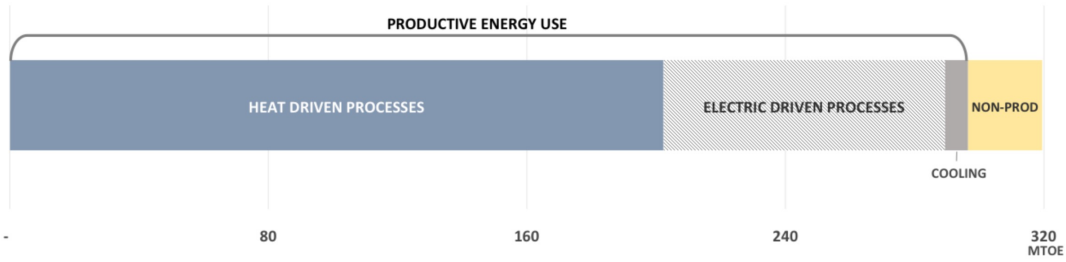
An EU-28 wide proxy for capturing demand from this source can be a consumption weighted average of regional values. However, since wholesale gas pricing varies regionally, we ought to only include those countries where gas is predominantly priced in relation to the hubs. As of 2015, according to Stern & Rogers (2014), NWE and CEE, in that order, are the only regions where hub pricing accounts for more than *50%* of the sales. They report a *15%* figure for the Italy, however, since this market is sizeable, we include it in our calculations as well.







Industrial sector has been since 1991 at times the second largest and at other times the third largest consumer of natural gas accounting for roughly one third of the total consumption (see Error: Reference source not found). Approximately two third of the energy consumed in this sector goes to heat driven processes integral to the production of output (see Error: Reference source not found). Majority of the energy for such functions, whereby temperature range of 100! C up to *1650 ! C* is required, is sourced from natural gas. In this regard chemical sector is the most energy intensive making up *30%* of industrial demand for natural gas (Pisca,



2017).

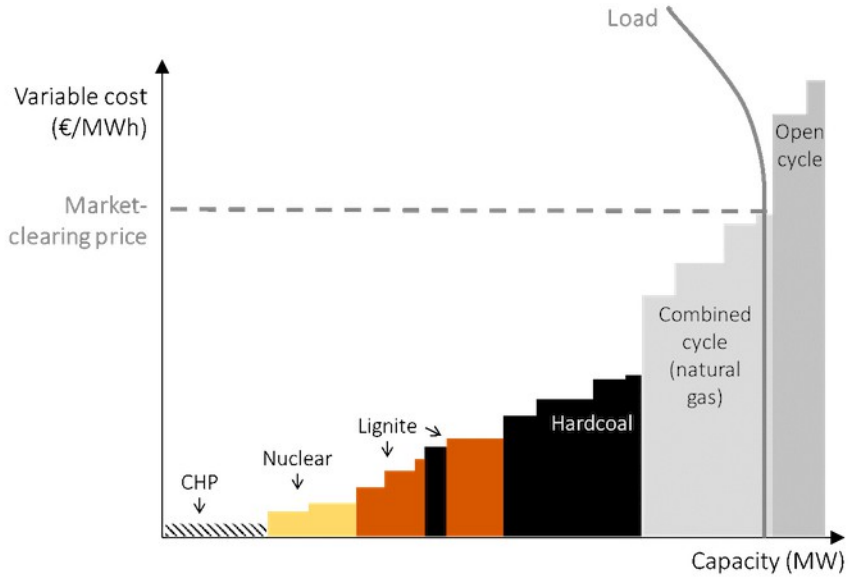
As the demand for energy from the industry is a function of their economic activity i.e. level of output, it makes sense to assume a degree of correlation between demand for gas and business cycles. In recession, therefore, lower gas prices and in growth phase higher prices are expected. Hulshof et al. (2016) uses a gas consumption weighted average of Industrial Production Index (*IPI*), compiled by Eurostat, in west Europe as a macroeconomic variable in their model. This index, which is calculated for all EU-28 on monthly basis with two months lag, measures monthly changes in the price-adjusted output of industry. In Error: Reference source not found we can see the seasonality and a certain degree of covariance between industrial activity of mining and manufacturing sectors and gas consumption. The trough in both timeseries coincides with July-October period; whereas, the peak in the industrial activity during the holiday season in December and January is less pronounced. The argument can be made that most of the industrial consumption is hedged through long-term contracts and only a minor fraction is sourced from the spot market for optimization purposes. However, since such type of data, i.e. the hedging ratio, is proprietary and difficult to obtain we proceed to include IPI as an independent variable in our model to capture the effect of wholesale industrial demand on the spot balance.

5.2. Power generation

An equally significant portion of the demand comes from the power sector (see Error: Reference source not found), whereby powerplants convert heating energy into electricity. The bidding volume for natural gas coming from this sector is, therefore, a function of generation capacity of the gas-fired powerplant fleet and the respective number of operation hours.

The European electricity market is a rather transparent market where data pertaining installed production capacity of different fuel types including natural gas in each regional market is publicized and retrievable from the transparency platform of

European Network of Transmission System Operators for Electricity (*ENTSOE*). Utilization rate of the available capacity is, though, a question of cost effectiveness.



Compared to lignite and coal powerplants, gas powerplants run mostly during the peak load hours where electricity prices are the highest. In a typical European liberalized electricity spot market, the day-ahead electricity prices for each hour, known as market clearing price (*MCP*) is determined through intersection of demand and supply curves. These curves are constructed from the bid and ask price-volume pairs submitted by, in turn, the consumers and producers. The supply curve is also called merit order stack as it represents the stacking of offered volume by the powerplants in ascending order of their running cost. Powerplants who offered volumes at prices lower or equal to the later calculated market clearing price are committed to run. Gas powerplants are more expensive to run, therefore, they rank higher in the stack. As a result, they run only in hours where MCP is higher than their running cost; and these are typically high load hours (see Error: Reference source not found). The primary cause behind higher running cost is higher commodity of price of natural gas in comparison to lignite or coal for every megawatt hour produced. Introduction of carbon tax, however, has turned gas more competitive as it emits *50%* to *60 %* less carbon dioxide compared to coal and lignite for same amount of energy produced.

Coal switching price is a theoretical threshold defined for a gas power plant with a certain efficiency at gas prices below which gas plants become more competitive than the coal-fired peers. This price is calculated using coal, gas, and emission prices in Formula 7:

(7)

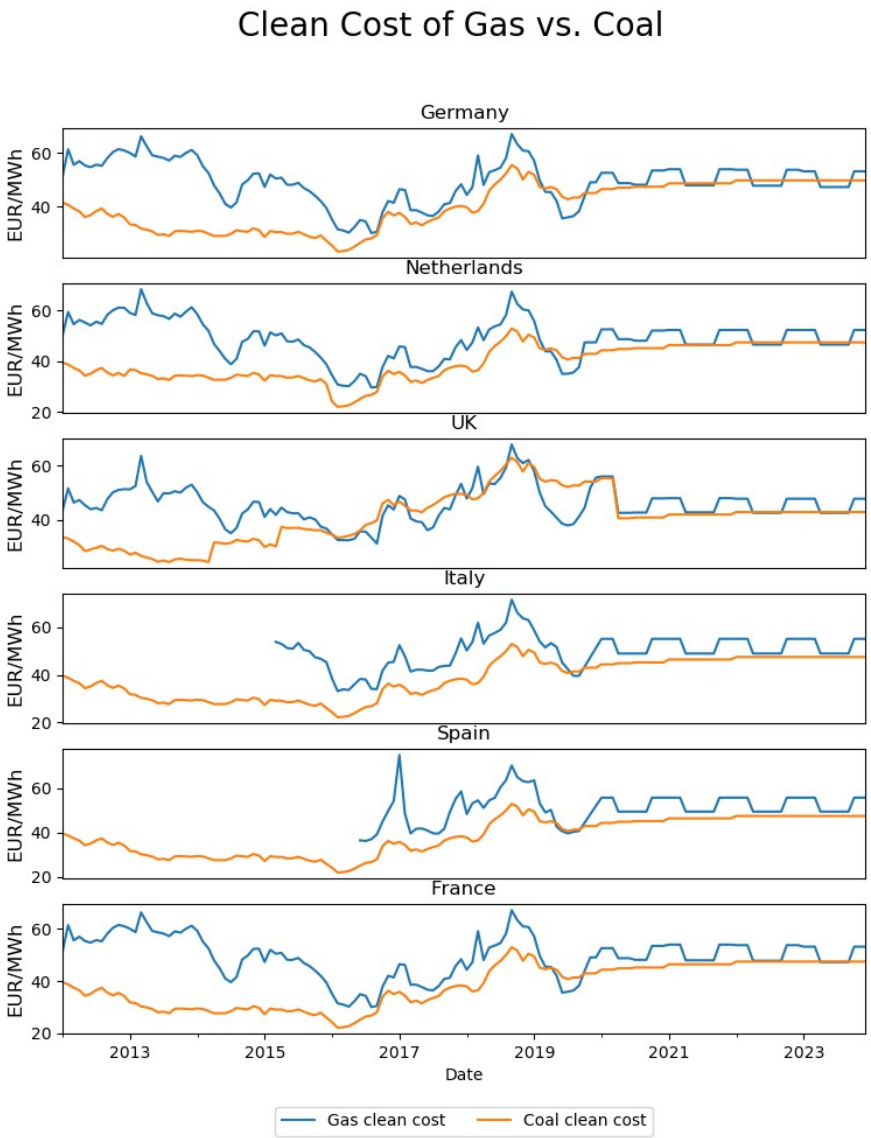
Where is the efficiency of a given gas powerplant e.g. *45% - 50%,* is the efficiency of a given coal powerplant e.g. *35%*, is the emission factor of a coal powerplant, is the emission factor of a gas powerplant, is the price of carbon emissions per metric ton, is the local coal tax and is the local carbon price support that only applies to the United Kingdom.

Emission tax in EU is settled through a financial product called European Emission Allowances (EUA) which is traded in both spot and forward market for different delivery months. Another way of understanding this concept is to calculate the clean cost of producing *1* MWh worth of electricity with gas and coal and then compare which is more cost-effective. Clean cost of production refers to the cumulative fuel, emission, and regionally applicable coal tax costs. Assuming constant efficiencies for both technologies, a competitive gas environment is achieved through lower gas and higher *EUA* and coal prices. In this environment, gas plants displace coal plants in the merit order stack and subsequently run more hours during the day. Moreover, the uncompetitive lower efficiency gas plants, that were previously rarely deployed, are more likely to operate during peak load hours. The net effect of both of these consequences translates to higher demand for gas.

Calculation of historical and forward clean cost of gas and coal shows gas as the more competitive fuel between January and September of 2019 (Error: Reference source not found). Unprecedentedly, in the six key markets i.e. UK, Italy, Spain, Germany, Netherlands and France that account for more than *70%* of the gas-fired generation capacity in Europe (Timera Energy, 2016), by virtue of cheap gas and high emission prices, gas has displaced coal in the merit stack for the first nine months of 2019 except for Italy where the displacement extends over a shorter span. This means for these periods of time gas turbines will be supplying the base load instead of just the peak load. It also translates to a diminishing idle capacity which means in case of a supply glut power sector will no longer capable of absorbing the bearish shock.

Another factor driving the demand from the sector is the amount of renewable power in the generation mix. Despite being capital intensive investments, renewable energy sources such as wind, solar, or hydroelectric power are associated with low variable cost. The more power generated from these sources the less is demand for fossil fuels.

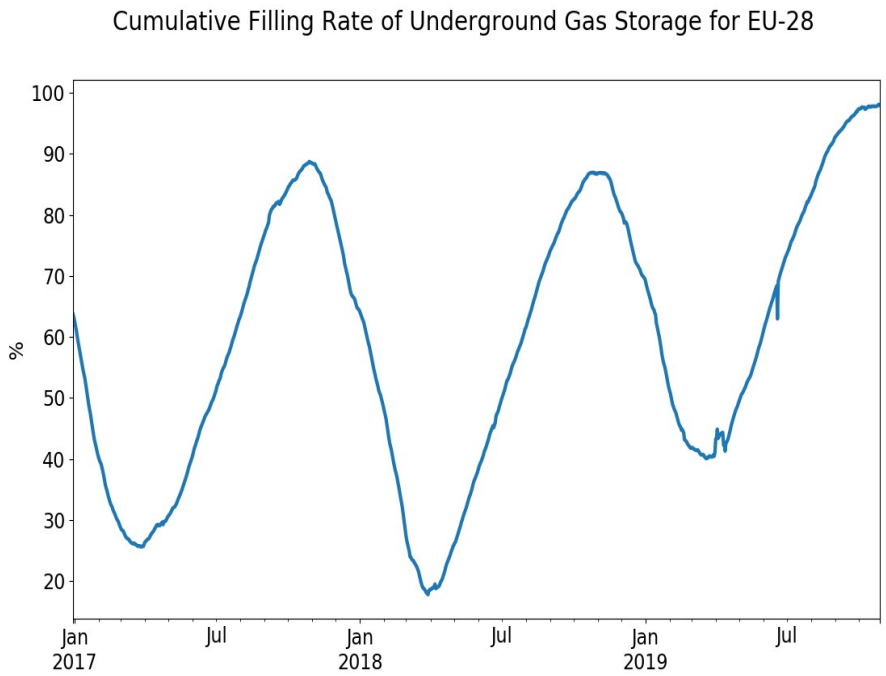
Low running cost of the renewable power decreases demand for hydrocarbon-based fuel by displacing them in the merit stack. In general, the long-run marginal cost (*LRMC*) of a wind or photovoltaic (*PV*) farm per unit of generation is less than that of a lignite plant, placing it next to nuclear or combined heat and power (*CHP*) powerplants (see Error: Reference source not found). As a result, the supply curve shifts to the right and for the same amount of load a lower MCP is achieved, at which a part or the entire gas-fired fleet is not economically viable to operate. As the renewable generation grows the supply curve shifts more towards the right bringing the MCP to lower levels.



In order to account for the negative correlation between renewable power generation and gas prices Nguyen & Nabney (2010) included wind speed as an exogenous variable in their short term price forecasting model using wavelet transforms. In their preprocessing stage they opted for including this variable as it contributed to better

results. Hulshof et al. (2016), on the other hand, used the day-ahead forecast of wind power generation for Germany to forecast TTF hub prices. Contrary to their pre-supposition, the sign of the variable was positive and significant at *1%*. The authors relate this observation to the "loop flow" phenomenon. Whereby in high-wind days in Germany the German-Dutch power interconnectors are scheduled to transmit less commercial flow in order to accommodate the so-called unplanned loop flows to avoid grid overload. Reduction of cheap imported renewable electricity from the German side to the Netherlands, therefore, leads to a steeper supply stack where more gas plants have to run; hence higher demand for gas.

5.3. Storage demand



In its transition towards a profit-based business, the gas storage sector is becoming a short-term demand and supply force. Nowadays in Europe, two types of gas storage facilities exist: 1. Underground Gas Storage (*UGS*) and 2. LNG storage facilities. LNG storages are more relevant to the supply side. This is due to the costly nature of liquefaction for the purpose of re-injection. In this section we focus on UGS. Europe's traditional approach to UGS has been that of supply security during high demand season namely winter; a cost-effective solution to the seasonality of consumption in comparison to the expansion of the production and transmission infrastructure. This means without the underground storage of the natural gas the seasonal price difference would be much higher. Stopa & Kosowski (2018) point out to the problem of financing new UGS investments in the current market environment where the seasonal spread is sliding below the average storage cost. According to

them, under such market conditions, new projects are approved with an eye on short-term price arbitrage opportunities rather than seasonal spreads.

Optimization operations of UGS owners can affect the daily balances in the gas spot market. As pointed out in the previous section the low spread environment in the recent years has given rise to storage owners interested in short-term speculation. Even though, as of 2019 the trend has reversed, and the winter-summer spread has more than normalized, let us for the sake of the argument suppose there are players that are speculative in nature as opposed to the traditional storage owner whose main strategy is materializing the value in the seasonal spreads by hedging their asset in a planned manner. Let us also assume that the speculative players cancel each other out and their activity has a net zero impact on the balance of the market. Then we can shift our focus to the more predictable players whose asset is the storage capacity and their main goal is to warehouse cheap summer delivery gas and sell it at profit in high demand winter period. This is quite evident in Error: Reference source not found, where storage facilities are at full capacity at the beginning Q4 and in effect empty by the end of Q1.

In order to understand the execution aspect of this strategy, one should know two conventions in the gas market and their ramifications. Firstly, the so called 'gas year' starts in October and ends in September; whereas, a gas storage contract year typically starts in April and ends in March (see Innogy, 2019 for a specimen of an annual gas storage contract). Secondly a gas year is divided into four quarters. In the second and third quarters (Q2 and Q3), also known as summer season, where prices are low due to low demand, the storage owner buys gas to store and subsequently withdraw to sell at a higher price in Q4 and Q1, also known as the winter season. Put differently, in summer gas gets injected and in winter withdrawn. This means full storage by beginning of October and empty by end of March. The injection and withdrawals are usually mostly hedged and follow a shallow trajectory path where the volumes are bigger in the beginning and wind down progressively. However, this monotonic behavior might be disturbed by DA-October and DA-March futures price spreads respectively in summer and in winter. For instance, if DA-October spread is positive, which mean day-ahead delivery is more expensive than October delivery, the operator pauses day-ahead planned injection and instead postpones it to October through selling the spread. In another scenario, if DA-March spread is positive, the operator is better of selling the planned withdrawal amount for March in the day-ahead market. It is important to note that these activities are not market drivers but rather offsetting in nature. In a tight balance set up, for example, where day-ahead contracts are trading at premium to October delivery, pausing injection is a

reactionary measure that instead of contributing to the tightness counterbalances it by shifting the demand curve to left. In short, UGS's short term optimization provide floor and ceiling for the price action and in this way, they affect the daily balance of the gas market. The magnitude of their impact as discussed can be proxied by the certain futures price spreads.

5.4. Demand shocks

Demand shocks are outlier events that tighten the balance of the market by a high degree of magnitude through creating an unexpected spike in demand. Depending on its longevity, a demand shock sends shock wave across the forward curve starting from the spot market all the way to the back end of the curve where the yearly products are. Because of their non-linear effect on the price, majority of regression analysis in the literature treats shocks as stand-alone dummy variables. In this paper we also follow the same tradition. In the next paragraphs, we will go over two of the well-known demand shocks in the recent history of the gas market.

In 2011, a tsunami ensuing an earthquake led to the malfunction of the cooling mechanism and subsequently meltdown of the nuclear reactors at Fukushima Daiichi powerplant in Japan. A year later, between 2012 and 2013, as a matter of safety revision, the entire Japanese nuclear fleet was suspended. At the time this developed industrial country produced *30%* of its electricity from nuclear reactors. Japanese power gird decided to balance itself by employing gas-fired powerplants. The utilities, therefore, had to import large amounts of LNG which led to a significantly tight balance in the gas market. Hulshof et al. (2016) proxy for this event using a two-state dummy variable to separate the periods before and after the event. Their analysis, however, did not find a significant impact on TTF spot prices.

In Europe the usual source of demand shocks are extreme sub-zero temperatures. During these temperatures, gas grid comes under a lot of pressure as the primary source of energy for space heating (refer to section Heating demand). In such circumstance, it is common for the grid operator to shut off the flow to the industrial sector to meet the residential sector demand. Hulshof et al. (2016) use a dummy variable to mark the 2012 cold spell period in Europe. For which also no significance impact on TTF price was found. In this paper, our methodology follows in the same footsteps i.e. accounting for each separate source of exogenous demand shock by introducing a two-state dummies into the model definition.

6. Supply

6.1. Imports

In contrast to demand, supply quantities to the continental Europe and EU in particular are relatively transparent. In that they are monitored, recorded and accessible to public on the transparency platform of the European Network of Transmission System Operators for Gas (ENTSOG). Therefore, in order to account for the total inflow of gas into the grid systems all is required is to sum up the net flows at destination hubs, in case of pipelines, and from terminals, in case of LNG, whereby supply flows arrive at.

6.2. Pipeline flows

Timera Energy (2013) categorizes the supplies into EU into four sources: Norwegian pipeline, Russian pipeline, North African pipeline and LNG imports. Norwegian supplies flow into NBP, TTF, and NCG hubs through a multitude of interconnections. Russian flows through two major pipelines Yamal and Nord Stream arrive in 10 destinations, from which German hubs of Gaspool and NCG accounts for more than half and Slovakian Virtual Trading Point (Slovakia VTP) in the second place attracts around *13%* (Sharples & Henderson, 2019). In the time of writing, Gazprom is the sole entity marketing Russian pipeline natural gas to EU with a total sales volume of 148 Bcm in 2018, almost one third of the EU's total consumption. North African pipelines flow only into Italy and Spain. Majority of these volumes are oil-indexed. The Spain bound volumes have no access to any major hub; whereas in Italy's case, they arrive at the Italian hub PSV where the possibility of arbitrage with the Northern hubs in low oil price environment exists. As of 2013, the annual import from North Africa stood at 40 Bcm (Timera Energy, 2013).

6.3. LNG imports

By the end of 2017, EU possessed annual LNG import capacity of 227 Bcm that covers roughly *40%* of its demand albeit with an average utilization rate of *25%*. Spain with 69 Bcm, UK with 43 Bcm and France with 35 Bcm are the largest importers (King & Spalding, 2018). The supply share of the LNG imports can be

calculated by summing up the send-outs from regassification terminals, which in the case of continental Europe is published on Aggregated LNG Storage Inventory (ALSI). As far as UK imported LNG is concerned, the interconnections connecting the island to the mainland Europe works as an aggregator on their own mixing LNG and production gas into one stream; therefore, it is not necessary to separately account for those volumes.

6.4. Storage

By the virtue of injection and withdrawal functionality, UGS can be a source of both supply and demand. Viewed as a supply source, withdrawal quantities are matter of public record accessible on Aggregated Gas Storage Inventory (AGSI) transparency platform for the EU countries. As discussed previously (see section Storage demandStorage demand), UGS operator has incentive to optimize their asset by taking advantage of price spread between day-ahead and March contract in winter and day-ahead and October contract in summer. From a theoretical standpoint, we can use this logic to predict UGS injection (demand) or withdrawal (supply) quantities. In general, since in terms of both demand and supply, the storages account for high quantities of gas, it is crucial to include the volumes coming in and out of the storing facilities.

7. Methodology

7.1. Description of data

The dataset comprises the variables described in Table 7 1. It spans from December 2017 to October. 2019. Although not all the original variable frequencies were homogenous, the dataset is resampled to daily granularity with the base hour of 6:00 Central European Time Zone as a matter of convention defined for a gas day. The base hour aggregation only matter in the case of hourly to daily computations. Furthermore, the seasonal, i.e. annual and weekly, differential of all non-spot-price variables are added to the dataset. The differential variables are denoted with suffixes '_d_7' in case of weekly and '_d_365' in case of yearly seasonality. In addition to that, lags of order one to seven of hub prices, including TTF, are also added to the input list. Lagged prices, on the other hand, have '_l_x' suffixes, where 'x' is can vary from 1 to 7. For instance, 'nl-ttf_l_1' stands for first lag of TTF spot prices. Additionally, a layer of geographical aggregation is applied to the data, whereby each fundamental variable's name is prefixed with the shorthand that refers to a certain region denoted as 'XXX' in Table 7 1. 'CWE_p_load', for instance, refers to power consumption in Central Western Europe.

In doing so, each region's data is either a consumption weighted aggregation, i.e. temperature, or total sum, i.e. power load, of its member states. The regions are as follows: EU-28, EU-27, Central Western Europe (CWE), Central Eastern Europe (CEE), Southeastern Europe (SEE), Scandinavia (NP), Iberia (IB), Baltic Countries (BLT), Great Britain (GB), and Ukraine (UA). The shorthand for these regions is borrowed from the nomenclature native to the European power markets. The member states of each region are listed in the appendix. The rationale behind such aggregation is dimensionality reduction. However, the reader might wonder if this particular set of bundles is necessarily economically sensible in application to Gas markets. To put this concern to rest, it should be stated that member states of each regions more or less mirror similar gas market profiles. For instance, in most of NWE countries, 80% of the price formation mechanism comes from gas competition and the rest comes from oil-linked long-term contracts; whereas, in SEE countries this ratios are reversed (Stern & Rogers, 2014). As another example, gas grids in both Iberian states, Spain and Portugal, are minimally integrated into the rest of Europe's gas system. Portugal's only access route to the rest of Europe is through Spain, and

Spain's only access route is a combined 225 GWh/d interconnection capacity to the South of France (Entsog, 2019). At full capacity this interconnection can only supply *25%* of Spain daily consumption.

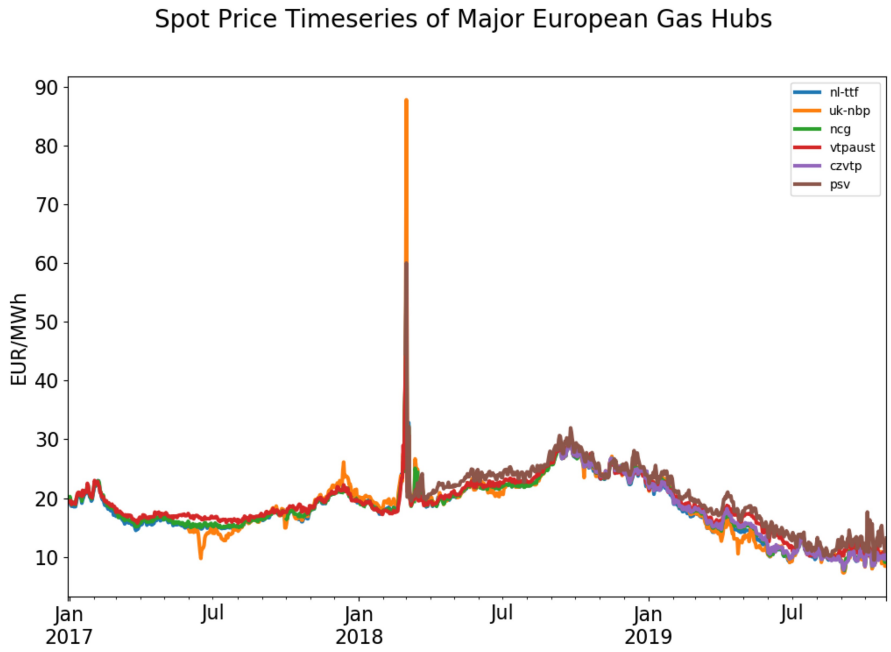
Table 71 Variable names and specifications

Variable Name	Description
XXX_P_load	*Power load. Aggregated geographically by summing up each member's power consumption.*
XXX_Wind	*Wind generated power. Aggregated geographically by summing up each member's wind power generation.*
XXX_Solar	*Solar generated power. Aggregated geographically by summing up each member's solar power generation.*
XXX_Temp_HD	*Heating degree. Defined as difference between actual temperature and standard Heating degree threshold, 15-degree Celsius, aggregated geographically by taking consumption weighted average of the member states.*
XXX_Temp_CD	*Cooling degree. Defined as difference between actual temperature and standard cooling degree threshold, 25-degree Celsius, aggregated geographically by taking consumption weighted average of the member states.*
XXX_Temp	*Temperature in degree Celsius, aggregated geographically by taking consumption weighted average of the member states.*
IP_EU_19	*Industrial Production Index calculated for 19 Euro area countries compiled by Eurostat on monthly basis.*
XXX_full_ugs	*Filling percentage rate of an underground gas storage facilities in respect to its full technical capacity. Aggregated geographically by summing up each member's portfolio of storage units.*
XXX_storage_ugs	*Stock level of underground gas storage facilities (TWh). Aggregated geographically by summing up each*

	member's portfolio of storage units.
XXX_injection_ugs	*Injection rate of underground gas storage facilities (GWh/d). Aggregated geographically by summing up each member's portfolio of storage units.*
XXX_withdrawal_ugs	*Withdrawal rate of underground gas storage facilities (GWh/d). Aggregated geographically by summing up each member's portfolio of storage units.*
XXX_inventory_lso	*Inventory level of LNG storage facilities (10^3 x m^3). Aggregated geographically by summing up each member's portfolio of storage units.*
XXX_sendout_lso	*Send out rate of LNG storage facilities (GWh/d). Aggregated geographically by summing up each member's portfolio of storage units.*
XXX_RU_IMP	*Total Russian pipeline import to XXX region (KWh/d)*
XXX_NO_IMP	*Total Norwegian pipeline import to XXX region (KWh/d)*
XXX_LY_IMP	*Total Libyan pipeline import to XXX region (KWh/d)*
XXX_AL_IMP	*Total Algerian pipeline import to XXX region (KWh/d)*
GB_GB_EXP	*Total pipeline export from continental Europe to Great Britain (KWh/d)*
GB_GB_IMP	*Total pipeline import from Great Britain to continental Europe (KWh/d)*
nl-ttf	*TTF hub spot price and nl-ttf_l_x is the xth lag*
uk-nbp	*NBP hub spot price and uk_nbp_ l_x is the xth lag*
czvtp	*CZVTP hub spot price and czvtp_ l_x is the xth lag*
cevtp	*CEGH hub spot price and cevtp_ l_x is the xth lag*
psv	*PSV hub spot price and psv_ l_x is the xth lag*

7.2. Spot gas hub prices

In this section, we are examining the time series characteristics of the endogenous variable namely TTF spot prices. We start by visualizing the time series along with other hub prices. In the Error: Reference source not found, the spot price time series of most prominent European gas hub are presented. A consistent feature of these graph is the premium of all hubs, prominently PSV hub, above TTF. This is in most part due to transmission fee from the most liquid delivery point, TTF, to other less-liquid hubs and bottlenecks during peak consumption periods. The spike in March 2018 coincided with unprecedented tightness in the market balances, whereby stocks were too low and temperatures were below normal. All prices refer to delivery prices settled at period for the delivery at period, which more specifically starts from 6:00

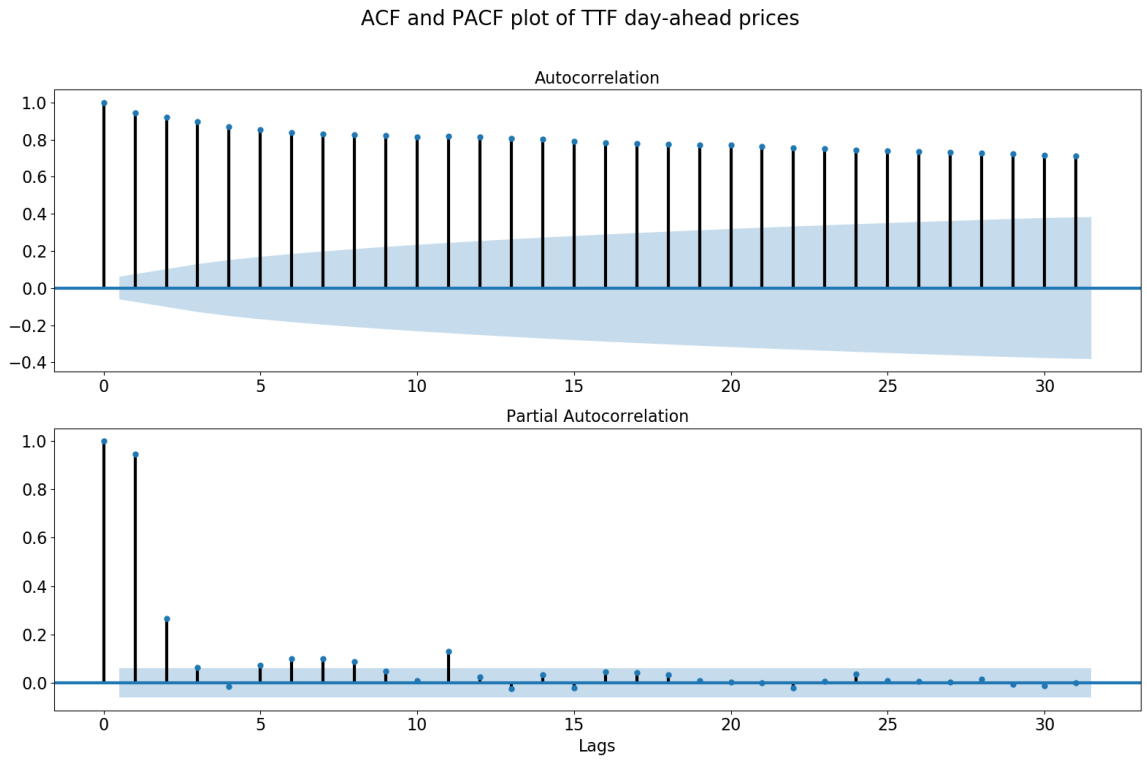


A.M. of day to 6:00 A.M. of day . The longest available set of observation is available for TTF and PSV hubs spanning from January 2017 to October 2019.

TTF spot prices is chosen to be the endogenous variable due its well-established status and unparalleled liquidity. Compared to other gas hubs, TTF has managed to overtake NBP and achieve the highest churn rate and most traded volumes in both spot and the futures market (Heather, 2012). Being the most liquid market and the market for which we have the largest amount of data at our disposal, it only makes sense to use this hub as the explained variable. Moreover, due to high degree of cointegration, verified by various studies, the results obtained for TTF will be more or less applicable to other gas hubs.

Results of the Dicky fuller test for TTF time series and its returns suggest, Table 7 2,

that nl-ttf is non-stationary; whereas, the p-value for its return is lower than critical value at *1%* confidence level and therefore the null hypothesis of non-stationarity can be rejected and the alternative hypothesis of stationarity can be accepted. In plotting the Autocorrelation and Partial Autocorrelation Functions, Error: Reference source not found, we observe firstly the momentum driven nature of the time series and secondly that the first two lags are the most significant. Later we see that they provide the most information gain.



able 72 Result of Dicky Fuller test on TTF spot prices and its returns

Timeseries	**p-value**	**Critical value at 1%**	**Critical value at 5%**	**Critical value at 10%**
nl-ttf	*-1.98*	*-3.4*	*-2.8*	*-2.5*
nl-ttf_d_1	*-17*	*-3.4*	*-2.8*	*-2.5*

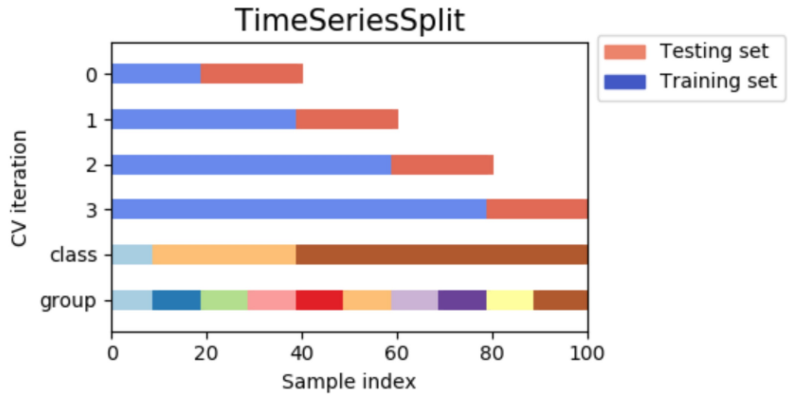
8. Methodology

In this section, the steps involved in the analysis process are described and discussed. These steps are listed and explained in chronological order starting with data collection and ending with estimation and performance evaluation.

The data used in this paper is sourced from various sources. A part of it is sourced directly from original sources through corresponding local APIs. For instance, the data related to underground gas and LNG storages were directly downloaded from corresponding transparency platforms i.e. Gas Infrastructure Europe. Similarly, data related to the physical pipeline flow was downloaded from transparency platform of European Network of Transmission System Operators for Gas. Gas consumption and Industrial Production Index data were sourced from Eurostat's database. The rest of the data was sourced from second hand data providers. This option was opted because of the time-consuming nature of the alternatives. For example, scraping power consumption data from each individual 27 European states' transmission system operator's API would have been timewise a very expensive option. The data sourced in this manner include power and weather-related variables as well as price timeseries. In the next step, the collected data needed to be preprocessed before fitting into the model.

The data processing in this work mostly entails aggregation, unit standardization and filtering. The aggregation process was two-faceted: geographical and temporal. For instance, the pipeline flow data is published for each transmission point individually rather than for bilateral flow between geopolitical borders. Therefore, the relevant transmission points had to be identified first and then aggregated for the corresponding geopolitical borders. In case of temporal aggregation, for instance, the hourly power and weather-related data had to be resampled to daily frequency by taking the mean of all 24 hours within a gas day. All data had to be resampled to daily granularity as the estimated variable spot prices used in our analysis is indeed the settlement price for the day ahead delivery. Moreover, in order to get rid of the bigger numbers, physical flow timeseries were converted to GWh/d from KWh/d. Storage data are by default reported in GWh/d. This conversion is not necessarily needed to arrive at correct estimations from our model, as decision tree algorithms are robust in relation to unscaled data. However, this was done to facilitate better comparison of magnitude of slope coefficients in the OLS regression analysis.

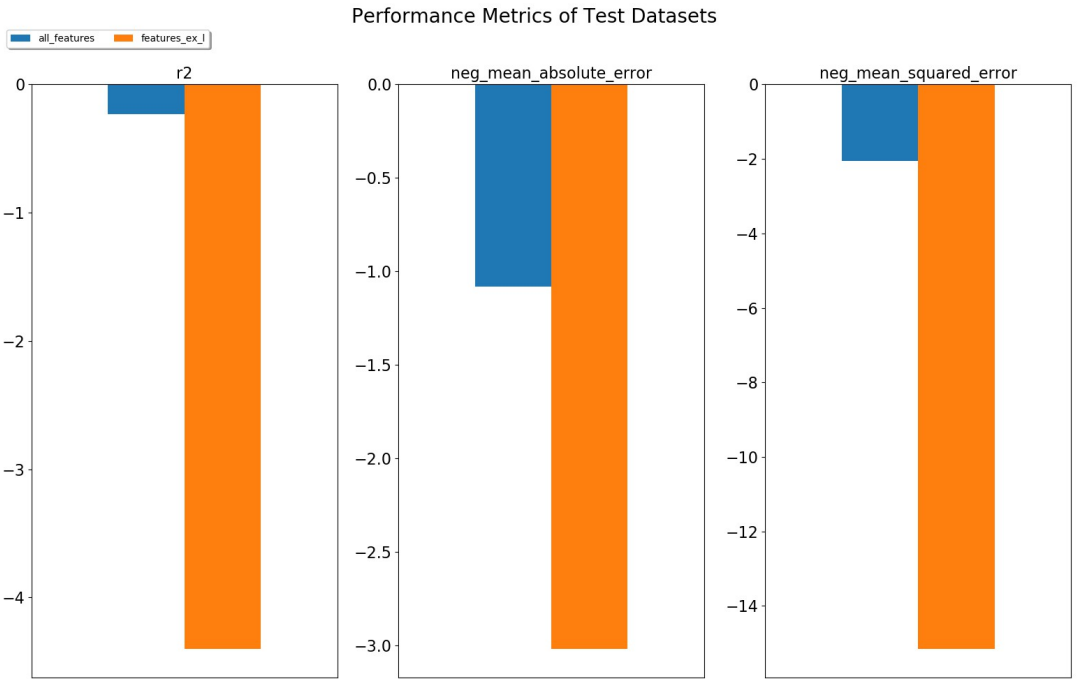
Finally, some of the timeseries downloaded from the data sources were not necessarily fundamentally relevant and did not have to be present in the final data fed into the model. Those unnecessary part of the data, therefore, needed to be filtered out, which also aided in cutting the computation time. At this point, the dataset is processed sufficiently for the model to be trained on.



In the estimation phase, the dataset is divided into train and test splits, whereby, the estimator is first trained on the 'train' dataset and then, in the second step, the trained estimator, with pre-fixed parameters, is used to predict the 'test' dataset. In machine learning estimation, Cross Validation (CV) is a technique whereby the model is initially parametrized to fit the training data and then validated on the test data. In case of a timeseries, in order to respect the chronological order in which the data is observed, the train sample is preceded by the train sample and they do not overlap. Following this rule, the data is then split in desired number of folds (see Error: Reference source not found). In a four-fold example, the data is split, and the model is trained, validated on the test sample in four iterations. The option for the train sample to either include all previous observations or only a maximum number of previous observations at each iteration is discretionary. In our analysis, we experimented with both number of folds and the number of maximum previous observations. In regard to setting the maximum training window, our analysis showed 90-day window yields the most optimal results. Having fixed this parameter, the estimations are then carried out on two different sets of data. The first dataset includes all variables including associated seasonal differentials and price-lags. Whereas, the second dataset, dubbed as 'features_ex_l', excludes price lags. At each iteration the predictions of both train and test data is then stored. Based on which error metrics are later calculated. In addition to error metrics calculation, the tree parameters are used to calculate different feature importance types mentioned in the theory section. In this paper, we use 'Percentage Gain' as a yardstick to measure the importance of each feature and its contribution to total information gain. Finally, after identifying the most important features, we run an OLS regression on those variables against the explained variable, nl-ttf, to understand the nature of the relationships and if they are significant at all.

9. Empirical results

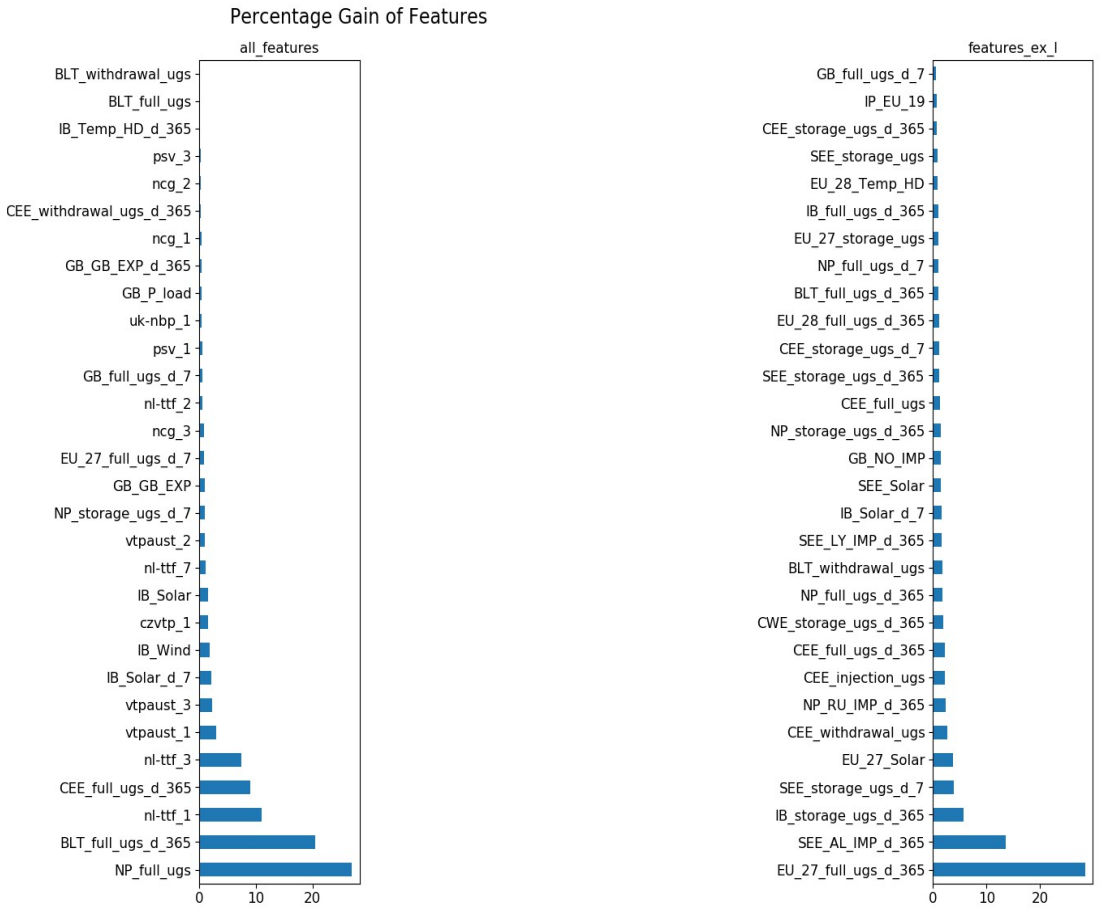
9.1. Model performance



n line with our expectations all the train dataset performance metrics outperformed the test counterparts. This stark difference arises from the fact that the forecasting tree is parametrized using the training dataset and naturally fits that set of observations much better. Contrasting the two models' test performances, as illustrated in Error: Reference source not found, results point out to the obvious superiority of the 'all_features' model, from this point onwards referred to as case A. This model's test error metrics both negative mean square error and negative mean absolute error are markedly lower than that of 'features_ex_l' model, from this point onwards referred to as case B. Similarly, in terms of goodness of fit, case A has a better R-squared score, whereas case B posts a markedly more negative one. The take away from this observation is as follows: inclusion of the lagged price and autoregressive components in the list of inputs significantly improves performance. Needless to say, contributions to better performance is not uniformly distributed. In the next section, we will identify features with the highest contribution for each individual case.

9.2. Feature importance

Percentage gain of each feature in both cases reveals that the contribution is concentrated among a handful of features (see Error: Reference source not found). In case A, more than *50%* of the total percentage gain comes from the NP_full_ugs, BLT_full_ugs_d_365 and first lag of nl-ttf. Similarly, in case B, with a an even higher degree of concentration, more that *50%* of total percentage gain comes, in turn, from EU_27_full_ugs_d_365, SEE_AL_IMP_d_365 and IB_storage_ugs_d_365. It is worth noting that, in general, seasonal differentials produce higher information gain than absolute value variables. Taking case B, SEE_storage_ugs_d_7, for example, has a *7%* percentage gain contribution; whereas, SEE_storage_ugs has less than *2%* contribution. The information gain of features is calculated from their appearance in the tree ensemble, and there can exist a high degree of correlation among them. When truncated by eliminating highly correlated features, we arrive at a slightly different yet more economically sensible list of features (see Error: Reference source not found). By removing high correlation

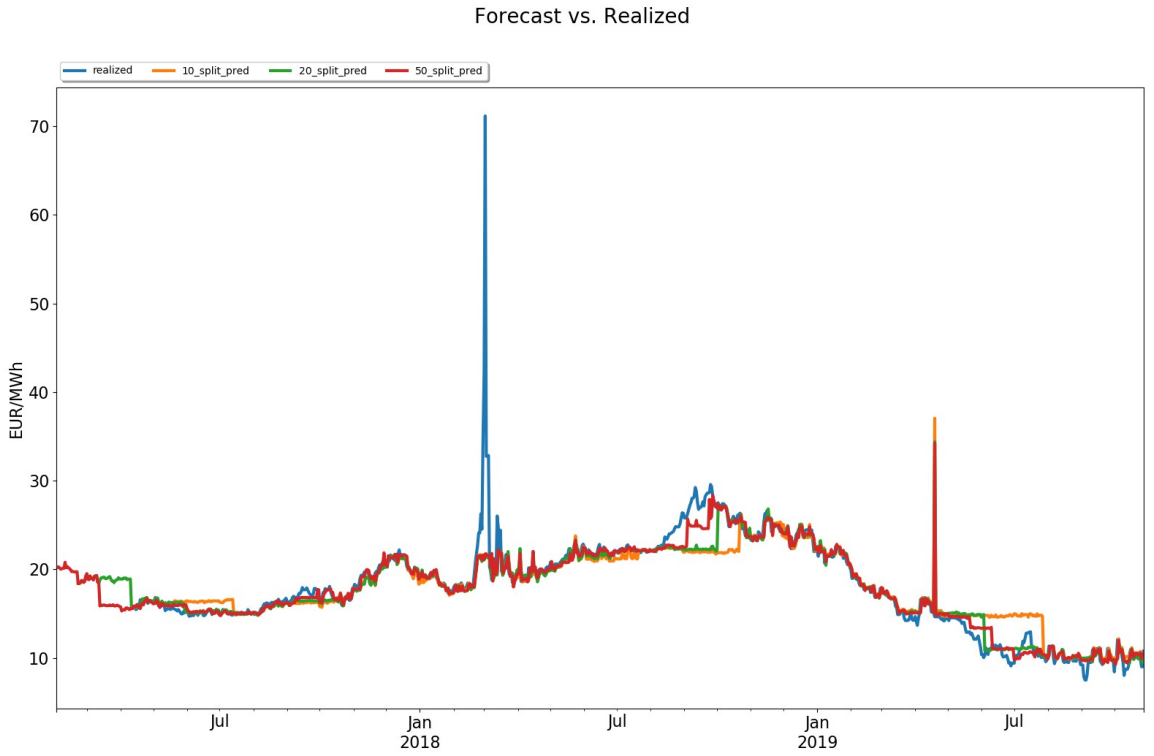


ariables in

case A, the distribution of information gain becomes even more concentrated with more than *90%* of the total gain belonging to top four features. These features, in order of importance, are as follows: BLT_full_ugs_d_365 (60%), nl-ttf_l (*18%*), CEE_full_ugs_d_365 (*8%*), IB_solar_d_7 (*5%*). Again, seasonally differenced variables rank higher than absolute value counterparts. Another interesting point here is how full percentage capacity of underground gas storages plays a consistently important role throughout. First degree autoregressive component, nl-ttf_l, is the second most valuable variable. This is due to autoregressive nature of TTF spot prices as discussed in the previous sections (refer to Error: Reference source not found). A less economically justifiable important feature is the Iberian solar power generation. Since Iberian gas grid is relatively less integrated to the rest of Europe, it is rather difficult to justify its impact.

9.3. Forecast and Discussion

By building a tree ensemble from low correlation variables, Error: Reference source not found, that make up *90%* of the total percentage gain, the algorithm produces the forecasts presented in Error: Reference source not found for various cross validation configurations.



here are three periods where the divergence between actual and forecast values are significant. The degree of divergence, however, subsides at higher cross-validation folds as the model learns to adapt to the new set-up more frequently. For instance, the divergence in the period June-August 2019 is more dramatic and prolonged in 10-fold split compared to 50-fold split. Notwithstanding the adaptability of the model at higher folds, the culprit for all the three occasions with high residual error lies in either unusually cold temperatures, unusually low storage levels or a combination of both. We go through each period to provide a more in-depth explanation.

During March-April 2018, period, the prices surged to highs of *70 EUR* and the estimator fail to capture the spike. From Error: Reference source not found, we observe significantly lower temperature compared to the same period the year before, at one point by -10 degrees. In addition to this, the end of March coincides with end of underground gas storage contract year where stock levels are bottoming out. Not only stock levels are low around this period but also in 2018, in particular, the levels were lower than previous year's levels by around *5%*, see Error: Reference source not found,. The unprecedent combination of these two bullish events led to unforeseeable spike in the spot market.

In an analogous manner, during the August-November 2018, period the forecast consistently underestimates the observed price. The undershooting is attributable on

one hand to UGS stock levels and on the other hand to low volatility training period preceding this interval where price is capped by *20-23* EUR range. In the same year, by the end of September, UGS stocks are expected to be entering the Winter Season with full capacity. However, one can see, as illustrated by Error: Reference source not found, that, contrasting to the previous year, the stocks were *5%* lower. This means the storage operators were still injecting well into end of November to reach full capacity. Therefore, we can conclude that the UGS demand in this period has had a supportive effect on the prices.

Unlike the last two episodes, the period May-July 2019, the forecast overshot the observed price. The divergence can be partially explained, as in the previous case, by the low volatility training data set, but the more influential factor lies in the UGS stock level. The unusually mild winter season in 2019 led to less demand. As a result, UGS stock levels, starting March 2019, were hovering around *10%* above previous year's level, Error: Reference source not found. In terms of supply and demand, higher than usual sock levels took injection demand away from the market catering for a bearish environment. The unprecedent positive annual stock difference by that point had not been observed in the training datasets. Consequently, the model did not possess appropriate memory to handle high surplus rate, hence reacted less dramatically. In conclusion, unseen and unprecedent seasonal differences in UGS stock levels and temperature have non-linear impact on the price action for prolonged period of time which is ultimately difficult for the ensemble of estimators to capture and extrapolate upon.

Results of linear regression of the top six features against the target variable, nl-ttf, indicates statistical significance of all features except IB_solar_d_7. Among the top most important features that contribute to more than 90% of total gain percentage, only weekly difference in Iberian solar power generation is statically insignificant. As previously pointed out, in the data description part, the first lag of the target variable is significant and has a positive coefficient hinting at momentum driven process. The rest of the high explanatory power variables pertain to seasonal difference of UGS stocks of various regions, refer to . The slope coefficient of which are all negative and significant at *1%* confidence interval. In practical terms, this means decrease in stock level compared to the previous period results in higher gas spot prices and vice versa. This is in line with economic expectation; in that, assuming everything else equal including demand from one period to the next, in the presence of lower quantities of supply, the system has to balance by bringing more expensive sellers of the commodity to the market, most probably storages with higher costs. Also, among the three most important regions, Nordpool weekly stock difference,

NP_storage_ugs_d_7, has the largest slope coefficient followed by the annual difference in storage filling rate of Central Europe, CEE_full_ugs_d_365, and Baltic, BLT_full_ugs_d_365, countries.

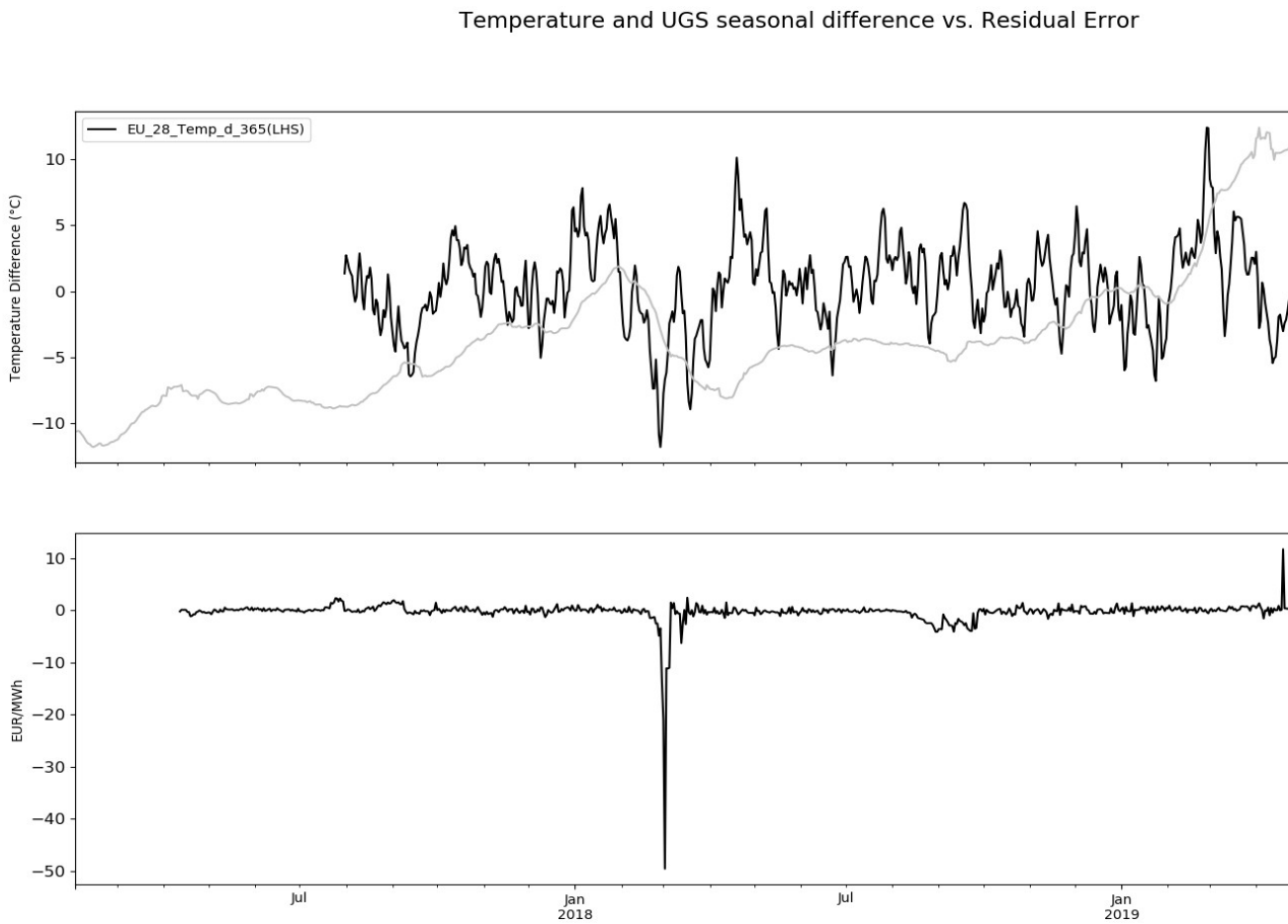


Table 93 Summary of results of OLS regression of top performing features on the target variable nl-ttf

	Slope Coef.	p-Value	Signficant at		
			1%	5%	10%
NP_full_ugs	*-0.132346*	*0*	*TRUE*	*TRUE*	*TRUE*
BLT_full_ugs_d_365	*-0.397829*	*0*	*TRUE*	*TRUE*	*TRUE*
nl-ttf_1	*0.998622*	*0*	*TRUE*	*TRUE*	*TRUE*
CEE_full_ugs_d_365	*-0.49339*	*0*	*TRUE*	*TRUE*	*TRUE*
IB_Solar_d_7	*0.000115*	*0.891807*	*FALSE*	*FALSE*	*FALSE*
IB_Wind	*-0.000084*	*0.232771*	*FALSE*	*FALSE*	*FALSE*
IB_Solar	*-0.006912*	*0*	*TRUE*	*TRUE*	*TRUE*
NP_storage_ugs_d_7	*3.231715*	*0.000817*	*TRUE*	*TRUE*	*TRUE*

GB_GB_EXP	*0.005803*	*0.000872*	*TRUE*	*TRUE*	*TRUE*
EU_27_full_ugs_d_7	*-0.223899*	*0.077538*	*FALSE*	*FALSE*	*TRUE*
GB_full_ugs_d_7	*-0.003838*	*0.89003*	*FALSE*	*FALSE*	*FALSE*
GB_P_load	*0.000187*	*0.004201*	*TRUE*	*TRUE*	*TRUE*
GB_GB_EXP_d_365	*-0.000715*	*0.467322*	*FALSE*	*FALSE*	*FALSE*
CEE_withdrawal_ugs_d_365	*0.003755*	*0.001404*	*TRUE*	*TRUE*	*TRUE*
IB_Temp_HD_d_365	*0.050175*	*0.685306*	*FALSE*	*FALSE*	*FALSE*
BLT_full_ugs	*-0.044991*	*0.000597*	*TRUE*	*TRUE*	*TRUE*
BLT_withdrawal_ugs	*0.010928*	*0.150191*	*FALSE*	*FALSE*	*FALSE*
BLT_P_load	*0.002095*	*0.029403*	*FALSE*	*TRUE*	*TRUE*
NP_storage_ugs_d_365	*-1.904459*	*0*	*TRUE*	*TRUE*	*TRUE*
NP_RU_IMP	*0.031995*	*0.001076*	*TRUE*	*TRUE*	*TRUE*
CWE_NO_IMP_d_365	*0.012096*	*0*	*TRUE*	*TRUE*	*TRUE*
IB_full_ugs	*-0.357278*	*0*	*TRUE*	*TRUE*	*TRUE*
SEE_Wind	*-0.000088*	*0.512337*	*FALSE*	*FALSE*	*FALSE*
GB_injection_ugs_d_365	*0.001464*	*0.390383*	*FALSE*	*FALSE*	*FALSE*
SEE_LY_IMP	*-0.028752*	*0.00032*	*TRUE*	*TRUE*	*TRUE*
CWE_inventory_lso	*-0.001385*	*0.179439*	*FALSE*	*FALSE*	*FALSE*
IB_full_ugs_d_365	*-0.480072*	*0*	*TRUE*	*TRUE*	*TRUE*
SEE_sendout_lso_d_365	*-0.014649*	*0*	*TRUE*	*TRUE*	*TRUE*

9.4. Seasonal difference in UGS stocks

With respect to the magnitude of slope coefficients and significance at various confidence intervals, presented in , there are number of variables that command markedly high explanatory power; hence, their economic sense need to be explored and verified. Among those variables, those that also contributed to the lion's share of the information gain in the tree ensemble, one can find filling rate of underground gas storages in the Baltic, Central Eastern Europe and Nordic countries. In addition to these fundamental inputs, TTF autoregressive component of order one is also quite significant. As discussed in the previous sections, spot gas prices can be categorized as momentum driven processes and as such lags and returns are expected to play a major role in determining the time series development. On the other hand, the aforementioned fundamental variables report peculiarly high significance in both estimations. Therefore, putting aside the question of causality, it is worthwhile to explain and illustrate the mechanics underlying these relationships. Among which, Baltic storage, in particular, reports unusually high level of information on top of

being statistically significant when regressed again TTF prices. Having a deeper look into that particular region will expose the reasoning behind the high correlation that more or less could be applied to other regions.

In the European context, Baltic underground storage is one of the least well-connected facilities and accounts for less than *2%* of the total capacity; yet, the yearly differential in the filling rate of the underground storages contribute to more than half of total percentage of information gain and significant with a coefficient of magnitude *-0.39* at *1%* confidence interval. In the next few paragraphs, we paint a picture of the Baltic gas market and illustrate the reason behind this phenomenon.

In terms of underground storages, Inčukalns UGS, located in Latvia is the only functional storage in the Baltic countries, which is supposed to ensure the regional security of supply. With *1.5* billion cubic meters per annum capacity, the storage is filled during the summer season, when consumption of natural gas in the region is several times lower than in the cold season, so that in the heating season it could be supplied to the customers in Latvia, Estonia, north-west Russia and, in smaller amounts, to Lithuania. Latvia has a unique, concentrated geological structure, which allows to create natural gas storages. Storage is possible because, in the deeps of Latvian soil there is a layer of porous sandstone, which has good storage properties and is coated with gas-tight rock layers. Also these geological structures are placed at optimal level of *700-800* meters deep, allowing safe and cost-efficient storage of gas (Conexus, 2019). Overall, it stands to reason that the volumes of this storage are nowhere enough to impact the balance of the European market.

Moreover, the region's only LNG import facility, named 'Independence', was built in Klaipeda port in 2014 to break the monopoly of Russia's Gazprom over gas supplies. Prior to 2017, it only imported Norwegian LNG for mostly domestic consumption, using between a fifth and a third of its annual capacity of *3.3* billion cubic meters, and began diversifying afterwards with the import of first U.S. cargo (Reuters, 2019) .

The region's transmission infrastructure can be described as an island. Even though, 'Alticconnector' is planned to connect Estonia and Finland by January 2020 and GIPL, Gas Interconnector Poland – Lithuania, is planned to be commissioned by end of 2021, at the time of writing of this paper, the Baltic region's transmission network is still under-developed with almost non-existent interconnection with the rest of European Union. On the other hand, the region is connected to Former Soviet bloc through a collective *13* Bcma set of interconnections that mainly transfer Russian gas directly or via Belarus. Russian gas enters the region through two main entry points namely Kotlovka and Korneti, in turn to Lithuanian and Latvian grids. The history of

capacity booking and physical flows on these two points reveals underutilization and abundance of uncontracted available capacity, Figure 9 1. (Entsog, 2019).

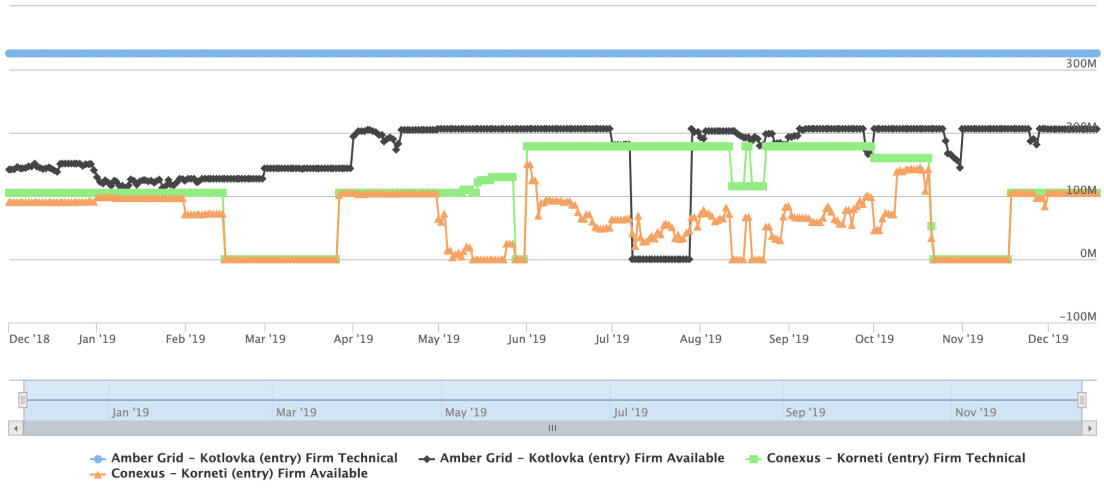


Figure 91 Firm technical and available capacity of Kotlovka and Korneti interconnection points in million KWh from December 2018 to December 2019

Source: (Entsog, 2019)

The high ratio of supply capacity to annual consumption of the region allows for sensitivity to higher spot prices given supply flexibility. According to the latest data, the annual gas consumption of the region stands at 5 Bcm (Eurostat, 2019). This amount can easily be covered multiple times over with the capacity of the infrastructure at hand. With a view to substantial uncontracted import capacity from Russia and a combined supply capacity of over 5 Bcm from LNG imports and underground storage already matching the domestic consumption, the region has the luxury of choice to optimize social welfare from various supply sources. In a day-ahead context, a Baltic agent can therefore opt to withdraw from storage in case Russian prompt volumes are too expensive. Assuming, Russian prompt volume are marketed at TTF hub price, the negative correlation between storage filling and TTF prices then starts to make sense. An alternative to Russian pipeline could be the uncontracted LNG stocks sitting at Klaipeda port marketed at a premium to TTF benchmark or somehow anchored to mainland spot prices. At any rate, the price sensitivity of storages accounts for the high information gain they produce in the estimation. In conclusion, the region is sensitive to spot prices and resorts to withdrawal from underground storage when faced with spikes in spot market.

The other two regions whose underground stocks play an important role in

determining the price are Central Eastern Europe and Scandinavia. Despite contributing less to the total information gain, yet their slopes are significant at *1%* confidence interval and in CEE's case even larger than that of the Baltic region. Unlike the Baltic region though, the combined storage capacity of the two regions is undoubtedly significant enough to render them as a price disrupter. Above and beyond, the interconnection infrastructure is in a developed stage. Therefore, it stands to reason that seasonally low stocks translate to higher short-term demand and subsequently translates to bullish price action and vice versa. So, while the results show that the stock levels in all the three regions are playing a role in setting the day-ahead prices, one should distinguish between actionary and reactionary forces. And in this context, the actionary regions are CEE and Scandinavia, to a lesser degree, and the reactionary region is the Baltic area.

10. Conclusion

In this paper, we estimated the day-ahead prices of the Dutch gas hub Title Transfer Facility for the very recent period of 2018-2019, in daily granularity, using Extreme Gradient Boosting Algorithm. This algorithm is a decision-tree based Machine Learning algorithm that uses a gradient boosting framework. It is both computationally efficient and algorithmically robust capable of handling high-dimensional datasets. On top of that, it is being employed by this paper for the first in the literature to forecast short-term gas prices. In doing so, numerous input variables were used that can be broken down into two categories: autoregressive and fundamental. Aside from autoregressive components, price lags of other European gas hubs were also part of the inputs. The cointegration among various European gas hubs is well reported in the literature. In an unprecedentedly comprehensive fashion, fundamental data pertaining to pipeline gas flows, underground gas and LNG storages, renewable energy generation, power consumption, and temperatures for all 28 European countries were collected and regionalized. The regionalized variables and corresponding seasonal differentials were then plugged in to two models: 1) one with all the inputs 2) one with fundamental variables only.

The estimation results of two configurations were in favor of including autoregressive component of order one in the model specification. As reported by other works in the literature, renewable energy generation appeared unimportant in both estimations. Moreover, we found out that yearly differential in underground gas stocks in Baltic, Central Eastern Europe and Scandinavia contain high amount information gain in our estimation. While the results show that the stock levels in all the three regions are playing a role in setting the day-ahead prices, one should distinguish between actionary and reactionary forces. And in this context, the actionary regions are Central Eastern Europe and Scandinavia, to a lesser degree, and the reactionary region is the Baltic area.

Moreover, the empirical results reported three periods where the model faced difficulty capturing the full swings of the price development. Notwithstanding the adaptability of the model at higher folds, the culprit for all the three occasions with high discrepancy resided in either unusually cold temperatures, unusually low storage levels or a combination of both. All in all, yearly differences in underground storage levels were found to be unequivocally decisive in the price development. This

correlation mostly manifest itself in non-linear fashion. As a topic of further investigation and as an extension to this paper, it would be interesting to examine the behavior of storage operators on micro level and model that behavior in relation to certain technical and market parameters. Modeling storage behavior helps us to move one step beyond understanding the history and instead forecast the future price development ex-ante which is the ultimate goal of this paper.

Bibliography

Čeperić, E., Zikovič, S., & Čeperić, V. (2017). Short-term forecasting of natural gas prices using machine learning and feature selection algorithms. *Energy*, *140*, 893–900.

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. Retrieved from https://arxiv.org/pdf/1603.02754.pdf

Conexus. (2019). Inčukalns UGS | Conexus. Retrieved December 21, 2019, from https://www.conexus.lv/incukalns-ugs-459

EC. (1998). *EC Energy Directive 98/30/EC*. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:31998L0030&from=EN

EC. (2009). Directive 2009/73/ec Of The European Parliament And Of The Council. *Official Journal Of The European Union*.

EEA. (2016). *Heating and cooling degree days*. European Environment Agency.

Entsog. (2019). Entsog—TP. Retrieved December 21, 2019, from https://transparency.entsog.eu/#/points/data?from=2018-12-01&indicators=Physical%20Flow%2CFirm%20Technical%2CFirm%20Booked%2CFirm%20Available&periodize=1&points=lv-tso-0001itp-00100entry

Eurostat. (2019). Eurostat—Data Explorer. Retrieved December 21, 2019, from https://appsso.eurostat.ec.europa.eu/nui/submitViewTableAction.do

Fattouh, B., Rogers, H. V., & Stewart, P. (2015). *The Us Shale Gas Revolution And Its Impact On Qatar's Position In Gas Markets*. Retrieved from Center On Global Energy Policy website: https://energypolicy.columbia.edu/sites/default/files/The%20US%20Shale%20Gas%20Revolution%20and%20Its%20Impact%20on%20Qatar's%20Position%20in%20Gas%20Markets_March%202015.pdf

Friedman, J., Breiman, L., Olsen, R. A., & Stone, C. J. (1983). Classification and Regression Trees. *Conference Proceedings*. Retrieved from https://www.semanticscholar.org/paper/Classification-and-Regression-Trees-Breiman-Friedman/8017699564136f93af21575810d557dba1ee6fc6

Geng, J.-B., Ji, Q., & Ying, F. (2016). The impact of the North American shale gas revolution on regional natural gas markets: Evidence from the regime-switching model. *Energy Policy*, *96*, 167–178.

Heather, P. (2012). *Continental European Gas Hubs: Are They Fit for Purpose?*

Hulshof, D., Maat, J.-P. der, & MachielMulder. (2016). Market fundamentals, competition and natural-gas prices. *Energy Policy*, *94*, Pages 480-491.

Innogy. (2019). *Specimen Yearly Gas Storage Contract with Firm Storage Capacity*. Retrieved from https://www.innogy-gasstorage.cz/media/smlouvy/Yearly_Firm_Capacity_Contract2020_sample.pdf

Jin, J., & Kim, J. (2015). Forecasting Natural Gas Prices Using Wavelets, Time Series, and Artificial Neural Networks. *PLoS ONE*, (10).

King & Spalding. (2018). *An Overview of LNG Import Terminals in Europe*. Retrieved from King & Spalding website: https://www.kslaw.com/attachments/000/006/010/original/LNG_in_Europe_2018_-_An_Overview_of_LNG_Import_Terminals_in_Europe.pdf?1530031152

Lochner, S. (2011). Modeling the European natural gas market during the 2009 RussianeUkrainian gas conflict: Ex-post simulation and analysis. *Journal of Natural Gas Science and Engineering*, *3*(1), 341–348.

Lochner, S., & Bothe, D. (2007). From Russia with Gas An analysis of the Nord Stream pipeline's impact on the European Gas Transmission System with the Tiger-Model. *EWI*.

Nguyen, H. T., & Nabney, I. T. (2010). Short-term electricity demand and gas price forecasts using wavelet transforms and adaptive models. *Energy*, *35*, 3674–3685.

Pisca, L. (2017). *European Union Industrial Energy Use With A Focus On Natural Gas*. Retrieved from Clingendael International Energy Programme website: https://www.clingendaelenergy.com/inc/upload/files/CIEP_2017__03_web.pdf

Platts. (2019). *Methodology and specifications guide*. Retrieved from https://www.spglobal.com/platts/plattscontent/_assets/_files/en/our-methodology/methodology-specifications/european_power_methodology.pdf

Regnard, N., & Zakoïan, J.-M. (2014). A conditionally heteroskedastic model with time-varying coefficients for daily gas spot prices. *Energy Economics*.

Reuters. (2019, January 21). Lithuania LNG port aims to be Baltic hub, double flows. *Reuters*. Retrieved from https://www.reuters.com/article/lithuania-lng-idUSL8N1ZL2C9

Schultz, E., & Swieringa, J. (2013). Price discovery in European natural gas markets. *Energy Policy*, *61*(C), 628–634. https://doi.org/10.1016/j.enpol.2013.06.0

Sharples, J., & Henderson, J. (2019). *Gazprom's Gas Sales via its Electronic Sales Platform (ESP)* (No. 51). Retrieved from OIES website: https://www.oxfordenergy.org/wpcms/wp-content/uploads/2019/07/Gazproms-Gas-Sales-via-its-Electronic-Sales-Platform-ESP-52.pdf?v=928568b84963

Stern, J., & Rogers, H. (2014). *The Dynamics of a Liberalised European Gas Market —Key determinants of hub prices, and roles and risks of major players*.

Stopa, J., & Kosowski, P. (2018). *Underground Gas Storage In Europe-energy Safety And Its Cost*.

Timera Energy. (2013, November). A framework for understanding European gas hub pricing. Retrieved from https://timera-energy.com/a-framework-for-understanding-european-gas-hub-pricing/

Timera Energy. (2016, May). Gas vs coal switching in Europe: Key markets [part I]. Retrieved from https://timera-energy.com/gas-vs-coal-switching-in-europe-key-markets/

Wattsight. (2019). Retrieved December 24, 2019, from https://wattsight.com/

Yorucu, V., & Bahramian, P. (2015). Price modelling of natural gas for the EU-12 countries: Evidence from panel cointegration. *Journal of Natural Gas Science and Engineering*, *24*, 464–472.

Appendix A: Regions

Regions	Shorthand	Member states
EU 28	EU_28	Germany, France, Austria, The Netherlands, Belgium, Sweden, Finland, Norway, Denmark, Poland, Czechia, Slovakia, Hungary, Italy, Romania, Bulgaria, Greece, Bosnia, Croatia, Macedonia, Montenegro, Serbia, Albania, Slovenia, Estonia, Portugal, Spain, Latvia, Lithuania, Great Britain
EU 27	EU_27	Germany, France, Austria, The Netherlands, Belgium, Sweden, Finland, Norway, Denmark, Poland, Czechia, Slovakia, Hungary, Italy, Romania, Bulgaria, Greece, Bosnia, Croatia, Macedonia, Montenegro, Serbia, Albania, Estonia, Portugal, Spain, Latvia, Lithuania
Central Western Europe	CWE	Germany, France, Austria, The Netherlands, Belgium
Central Eastern Europe	CEE	Poland, Czechia, Slovakia, Hungary
Southeastern Europe	SEE	Italy, Romania, Bulgaria, Greece, Bosnia, Croatia, Macedonia, Montenegro, Serbia, Albania, Slovenia
Iberia	IB	Spain, Portugal
Baltic region	BLT	Estonia, Latvia, Lithuania
Scandinavia	NP	Sweden, Norway, Finland, Denmark

Appendix B: Dataset (DVD)