Data Markets for Collaborative Forecasting in the Energy Sector

Tiago Guedes Teixeira
Master’s degree in Data Science
Computer Science Department
2023

Supervisor
Carla Gonçalves, Senior Researcher, INESC TEC

Co-supervisor
João Vinagre, Assistant Invited Professor, Faculty of Science, University of Porto

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

“ Ski-bi dibby dib yo da dub dub Yo da dub dub Ski-bi dibby dib yo da dub dub Yo da dub dub ”

Scatman

Acknowledgements

I would like to express my deepest gratitude to all those who have played a significant role not only in the process of crafting this dissertation but also throughout the entirety of my life's journey.

First and foremost, my heartfelt thanks go to my supervisor, Dr. Carla Gonçalves, whose unwavering patience, guidance, and cooperation proved instrumental in bringing this work to fruition. Her technical expertise and motivational support were pivotal in shaping this thesis. I am also thankfull to my co-supervisor Prof. Dr. João Vinagre.

I extend my sincere appreciation to my parents and my sister for their formidable support, encouragement, and belief in my abilities. To my mother Maria for her unconditional love; to my father Celestino for his daily support; to my swiftie sister Rita for her care and positivity.

A special thanks is reserved to a friend of mine, Vasco Tavares, who help me daily in this elegant fall into the unmagnificent lives of adults.

I also want to express my gratitude to lifelong friend Lucas Stein for his consistent encouragement and positivity throughout this academic journey.

To all my friends, family members, and colleagues who have offered their support, encouragement, and understanding during this process, I am immensely grateful.

Finally, I extend my heartfelt thanks to the entire academic community of University of Porto, whose collective knowledge and insights have shaped my educational experience.

UNIVERSIDADE DO PORTO

Abstract

Faculdade de Ciências da Universidade do Porto

Departamento de Ciência dos Computadores

Master's degree in Data Science

Data Markets for Collaborative Forecasting in the Energy Sector

by Tiago TEIXEIRA

The growing presence of geographically dispersed renewable energy sources, such as wind turbines, photovoltaic panels, and sensors, has generated extensive datasets that hold immense potential for advancing renewable energy forecasting. However, the reluctance of data owners to share their information, even with stringent privacy assurances, hinders collaborative efforts that could prove beneficial to all parties involved.

To address this challenge, this study introduces a novel data marketplace aimed at stimulating cooperation among diverse data owners. The marketplace employs a bidding mechanism that accommodates the needs of both data sellers and buyers, all while upholding data privacy. Sellers are empowered to specify their minimum compensation requirements for sharing data, while buyers indicate their willingness to pay based on anticipated improvements in forecasting accuracy.

At the core of this innovative framework lies a market operator tasked with data aggregation and forecasting, employing Splines Bid-Constrained Lasso Regression ($\mathcal{SBCLR}$). This approach effectively aligns the interests of all participants, resulting in mutually advantageous outcomes. Empirical evidence from various agents demonstrates the advantages of market participation, including enhanced forecasting skills and the potential for revenue generation through data sales.

Keywords: Renewable Energy Sources Forecasting, Data Markets, Bid-Constrained Regression, Data Sharing Incentives, Collaborative Learning

UNIVERSIDADE DO PORTO

Resumo

Faculdade de Ciências da Universidade do Porto

Departamento de Ciência dos Computadores

Mestrado em Ciência de Dados

Mercados de Dados para Previsão Colaborativa no Setor da Energia

por Tiago TEIXEIRA

O crescente número de fontes de energia renovável, como turbinas eólicas, painéis fotovoltaicos e sensores distribuídos geograficamente, tem gerado grandes volumes de dados com potencial para aprimorar as previsões de energia renovável. Contudo, a relutância dos proprietários de dados em compartilhar suas informações, mesmo com garantias rigorosas de privacidade, dificulta esforços colaborativos que poderiam ser benéficos para todas as partes envolvidas.

Para enfrentar esse desafio, este estudo apresenta uma nova plataforma de mercado de dados destinada a estimular a cooperação entre diversos proprietários de dados. A plataforma utiliza um mecanismo de licitação que atende às necessidades tanto dos vendedores de dados quanto dos compradores, tudo isso preservando a privacidade dos dados. Os vendedores têm a capacidade de especificar seus requisitos mínimos de compensação para compartilhar dados, enquanto os compradores indicam sua disposição para pagar com base nas melhorias na precisão das previsões.

No cerne deste inovador framework encontra-se um operador de mercado responsável pela agregação de dados e previsões, utilizando a Regressão de Lasso com Restrições de Licitação de Splines ($\mathcal{SBCLR}$). Essa abordagem alinha eficazmente os interesses de todos os participantes, resultando em resultados mutuamente vantajosos. Evidências empíricas de vários agentes demonstram as vantagens da participação no mercado, incluindo o aprimoramento das habilidades de previsão e o potencial de geração de receita por meio da venda de dados.

Contents

List of Figures

List of Tables

List of Algorithms

Glossary

RES	Renewable Energy Sources
MAE	Mean Absolute Error
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
LASSO	Least Absolute Shrinkage and Selection Operator
GBT	Gradient Boosting Tree
GP	Gaussian Process
AR	Auto-Regressive
MA	Moving Average
ARMA	Auto-Regressive Moving Average
ARIMA	Auto-Regressive Integrated Moving Average
$\boldsymbol{BGOT}$	Bid Gain Option Table
$\boldsymbol{\mathcal{SBCLR}}$	Splines Bid-Constrained Lasso Regression

Symbols

For notation purposes, vectors and matrices are denoted by bold lowercase and bold uppercase letters, e.g., $\mathbf{a}$ and $\mathbf{A}$, respectively. The vector $\mathbf{a} = [a_1, \ldots, a_k]^\top$ represents a column vector with k dimension, where a_i denotes scalars, $\forall i \in \{1, \ldots, k\}$. The column-wise joining of vectors and matrices is indicated by $[\mathbf{a}, \mathbf{b}]$ and $[\mathbf{A}, \mathbf{B}]$, respectively. Random variables are denoted by italic uppercase letters, e.g., Y. Estimators and estimates are denoted by hat operator "^", e.g., $\hat{Y}$ is the estimate for Y.

The main symbols are summarized here:

Notation	**Description**
N	Number of power agents
$\mathcal{A}$	Set of all agents
$Y_{i,t}$	Power measurements for RES agent i at time t
$Y_{i,t-l}$	l-lagged power measurements of agent i at time t
$X_{i,t}^{i'}$	i'^{th} variable from the i^{th} agent at time t, $X_{i,t}^{i'} \in \mathbb{R}^{n_i}$
n_i	Number of exogenous variables of agent i
$\mathcal{X}_{i,t}$	Local data from agent i at time t
$\boldsymbol{X_i}$	Local feature matrix containing historical data
H	Prediction horizon
T	Number of time steps in historical data
$\mathcal{Z}_t$	Data from all agents at time t
$\boldsymbol{Z}$	Feature matrix of historical data
$f_{\text{local}}^{(i)}$	Local model to forecasts of $Y_{i,t}$
$f_{\text{market}}^{(i)}$	Collaborative model to forecasts of $Y_{i,t}$
$\boldsymbol{\Theta_i}$	Collaborative model parameters to predict $Y_{i,t}$
$\boldsymbol{\beta_i}$	Collaborative model parameters of exogenous variables
$\boldsymbol{\eta_i}$	Collaborative model parameters of endogenous variables

$\boldsymbol{z_t}$	Observed values for $\mathcal{Z}_t$
$\boldsymbol{y_{i,t}}$	Observed values for $Y_{i,t}$
λ	LASSO regularization parameter
D	Splines order
K	Number of interior knots
$\tilde{\mathcal{Z}}_t$	$\mathcal{Z}_t$ after spline transformer
τ_i	i^{th} knot
$B_{i,d}$	d^{th} spline basis function for modeling $Y_{i,t}$
$\mathcal{T}(\cdot)$	Gradient Bosting Tree
$\mathcal{L}(\cdot)$	Loss function
$\boldsymbol{\theta}$	Bayesian Optimization parameters
$\boldsymbol{K}$	Kernel Matrix
$\mathcal{S}$	Set of sellers
$\mathcal{B}$	Set of buyers
p_i	Price to be paid by buyer i
$\boldsymbol{X_i}^{(val)}$	Explanatory validation set for $f_{\text{local}}^{(i)}$
$\boldsymbol{Z}^{(val)}$	Explanatory validation set for $f_{\text{market}}^{(i)}$
Δ	Size of validation set
$\mathcal{G}(\cdot)$	Gain function
b_i	Public bid of buyer i
$\mathcal{VF}_i(\cdot)$	Value function of buyer i
$s_{i,j}$	Seller i bid for the usage of their j^{th} variable
δ_b	Minimum non-zero sum between any pair of seller bids
$\tilde{\boldsymbol{b}}$	Potential bids in $BGOT$
$\mathcal{I}(\cdot)$	Indicator function
$CCF_{i,j}(\cdot)$	Cross-correlation function
$\mathcal{ST}$	Spline Transformer

Chapter 1

Introduction

1.1 Background and context

Renewable energy sources play a critical role in addressing climate change and achieving a sustainable future. The European Union has set ambitious targets to increase the share of renewable energy in final energy consumption to at least 32% by 2030 [1]. However, integrating renewable energy into the power grid efficiently poses challenges due to the intermittent nature and dependency on the weather conditions of these energy sources. Accurate forecasting of renewable energy generation is vital to ensure its reliability and optimize its integration, leading to enhanced efficiency and cost reduction. Recent research has demonstrated that combining data from different geographic locations can significantly improve forecasting models by capturing spatiotemporal dependencies [2]. Weather conditions at a particular location and time are not independent of their previous states, and neighboring locations exhibit correlated weather patterns. For instance, Zhu et al. proposed a unified framework for wind speed prediction that learns temporal and spatial correlations jointly. Their research demonstrated that incorporating spatio-temporal data improved the accuracy of wind speed forecasting. Cavalcante et al. explored the application of LASSO vector autoregression structures for very short-term wind power forecasting. Their study highlighted the advantage of incorporating spatial and temporal information to enhance the accuracy of wind power predictions. Similarly, the work [5] focused on solar power forecasting and demonstrated the effectiveness of sparse vector autoregression structures, which take into account both spatial and temporal relationships.

However, a significant barrier to achieve this data combination is the decentralized ownership of data, where multiple parties maintain their databases and are often unwilling to share due to personal or competitive reasons. Therefore, establishing incentive mechanisms for data sharing is crucial. Data privacy serves as a key incentive — for instance, training a linear regression model that combines data from multiple parties without revealing any private information [6]. However, relying solely on data privacy might not be sufficient. For example, data owner A can have huge improvements in accuracy when using data from data owner B, while data owner B does not have any benefit when using data from A. Although data privacy is ensured, why should data owner B cooperate?

To address this issue, a second family of incentive mechanisms has emerged: data markets, where the concept revolves around monetizing data. In fact, one effective approach to incentivize agents to share their data is by offering monetary compensation [7, 8]. One of the primary challenges in data markets is determining the value of data and establishing fair pricing mechanisms. The work in [9] was a pioneer in constructing a data market where buyers purchase forecasts instead of specific datasets based on cooperative game theory. Noteworthy features of this framework include equitable revenue distribution among sellers, payment based on improved forecasting skills, and compensation for incremental gain. However, this framework has limitations, including the inability to handle continuous updating of input variables, challenges with highly-correlated features, and the potential for agents to pay for redundant data.

To address these limitations, [10] presents an adapted framework for electricity markets that considers continuous updating of the input variables, incorporates lagged time-series data, and establishes a relationship between the exchanged data and reduction of imbalance costs. Nonetheless, this marketplace also has its own limitations, such as the lack of consideration for sellers' perspective and the computational complexity in the data monetization front, unsuitable for intra-day markets.

These limitations highlight the necessity for further refinement of marketplace design. Balancing the interests of both buyers and sellers is crucial, along with the development of a fast-running marketplace capable of handling short-term forecasting demands and real-time decision-making.

1.2 Research Questions and Contributions

This section outlines the central research questions addressed in this thesis and the corresponding contributions that constitute a slight advance the field of data markets, more oriented to RES Forecasting.

1.2.1 Research Questions

(**1**) **Establishing a Collaborative Data-Sharing Environment:** How can we create an environment for collaborative forecasting in data markets that allows sellers to specify their desired earnings and buyers to propose payments?

This research question delves into the exploration of mechanisms that serve the interests of both sellers and buyers while facilitating the expression of their requirements to participate effectively in the market. It aims to find a balance that accommodates diverse needs and objectives.

(**2**) **Integrating Constraints into Forecasting Methodology:** How can we effectively integrate the constraints determined by the interests of buyers and sellers into the forecasting methodology?

This question emphasizes the incorporation of seller and buyer requirements into the forecasting process while preserving the accuracy and reliability of forecasts.

1.2.2 Contributions

In the pursuit of addressing these research questions, this thesis makes the following contributions:

(**1**) **Exploring Regression Models in RES Forecasting:** We present an exploration of regression models in Renewable Energy Sources (RES) forecasting. Specifically, we demonstrate that Spline Lasso regression models can be employed effectively in RES forecasting. These models exhibit comparable prediction power to other model types while requiring less computational time. This contribution enhances the understanding of suitable models for RES forecasting.

(**2**) **Bidding Mechanism:** We introduce a novel bidding mechanism within our data market. This mechanism enables sellers to receive compensation aligned with their data-sharing requirements. Simultaneously, it empowers buyers to express their valuation of gain. This contribution facilitates a fair and efficient exchange of data in the market.

(3) **Forecasting Framework:** We implement a forecasting framework designed to handle the unique requirements of sellers and buyers in data markets. This framework effectively integrates constraints into the forecasting methodology, ensuring the market operates harmoniously while preserving forecasting accuracy. Furthermore, it incorporates a robust model that precludes the allocation of redundant features by the leverage of an integrated LASSO shrinkage method into our forecasting methodology.

1.3 Thesis Structure

The thesis is structered into 5 major chapters. A brief description of each one is provided.

Chapter 1 - Introduction: This initial chapter sets the stage for the entire thesis. It provides a concise introduction to the research domain, offering insights into the contextual background. It outlines the overarching goals and contributions of the research. Furthermore, this chapter introduces the reader to the organization and structure of the thesis, establishing a coherent roadmap for the subsequent chapters.

Chapter 2 - Problem Formulation and Literature Review: The second chapter formulates a standard RES forecasting problem. Additionally, it conducts a literature review, shedding light on the primary forecasting methodologies currently employed in the field. This chapter also provides an overview of existing solutions and approaches in collaborative forecasting, offering a valuable backdrop for the proposed research.

Chapter 3 - Proposal: Chapter three presents a comprehensive exploration of the proposed solution of a data market. It elaborates on the intricate mechanics of this data market and outlines the employed forecasting methodology.

Chapter 4 - Case Study: In this chapter, the research findings come to life through a real-world case study. The thesis presents a meticulous analysis based on data from ten wind farms in Australia. The case study serves a dual purpose: demonstrating the practical viability of the proposed data market concept and facilitating a comparative evaluation of various forecasting methods in the RES domain.

Chapter 5 - Conclusion: The concluding chapter discusses the research done and summarizes the contributions. Additionally, it provides future work to improve our proposal.

Chapter 2

Problem Formulation and Literature Review

The thesis aims to design an alternative data market for a day-ahead forecasting problem that involves multiple RES power plants. For this reason, we first introduce the notation and formulate the forecasting problem (Section 2.1). Then we review the main forecasting models (Section 2.2) and, finally, the more relevant data markets algorithms (Section 2.3).

2.1 Formulation of the RES Forecasting Problem

Consider a set of N agents, denoted as $\mathcal{A} = \{1, 2, \dots, N\}$, which represent the owners of the RES power plants. For simplicity, we assume that agent i operates a single power plant, having a single target variable $Y_{i,t}$, corresponding to the produced power. Additionally, each agent observes a set of n_i exogenous variables, such as wind speed or direction forecasts, denoted by $\mathbf{X}_{i,t} = \left\{X^1_{i,t}, \dots, X^{n_i}_{i,t}\right\}$. Here, $X^j_{i,t}$ denotes the j^{th} variable from the i^{th} agent, $i \in \mathcal{A}$.

At time t_0, the goal of agent i is to forecast $\{Y_{i,t}\}_{t=t_0+1}^{t_0+H}$, where H is the number of timestamps ahead. Two models can be considered depending on the available data:

- **Local model.** In a scenario where agent i only has local data, a model function $f(\cdot)$ can be used for each horizon such that

$$Y_{i,t} \approx f^{(i)}_{\text{local}}(\mathcal{X}_{i,t}; \Theta_i), \tag{2.1}$$

where Θ_i are the parameters to be estimated,

$$\mathcal{X}_{i,t} = \left\{ \underbrace{X^1_{i,t}, \ldots, X^{n_i}_{i,t}}_{\text{exogenous from agent i}}, \underbrace{Y_{i,t_0-1}, \ldots, Y_{i,t_0-L}}_{\text{L most recent power meas.}} \right\}. \tag{2.2}$$

- **Collaborative model.** In a scenario where agents share their data, a more robust function $f(\cdot)$ could be considered:

$$Y_{i,t} \approx f^{(i)}_{\text{market}}(\mathcal{Z}_t; \Theta_i), \tag{2.3}$$

where $\mathcal{Z}_t = \{\mathcal{X}_{1,t}, \mathcal{X}_{2,t}, \ldots, \mathcal{X}_{N,t}\}$, i.e.,

$$\mathcal{Z}_t = \left\{ \underbrace{X^1_{1,t}, \ldots, X^{n_1}_{1,t}, Y_{1,t_0-1}, \ldots, Y_{1,t_0-L}}_{\text{power plant 1}}, \ldots, \underbrace{X^1_{N,t}, \ldots, X^{n_N}_{N,t}, Y_{N,t_0-1}, \ldots, Y_{N,t_0-L}}_{\text{power plant N}} \right\}. \tag{2.4}$$

The mapping function $f(\cdot)$ can exhibit either linear or nonlinear characteristics, depending on the prediction horizon, and a multitude of approaches can be explored to model this function effectively. The forthcoming section will delve into a comprehensive discussion of the approaches to model $f(\cdot)$.

2.2 Forecasting Models

In the literature, there are two primary ways to classify methods used for forecasting renewable energy sources [11]: based on the prediction horizon or the methodology employed.

Regarding the prediction horizon, power forecasting can be categorized into four main groups based on the prediction time horizon, commonly referred to as ultra-short term (up to 6 hours), short-term (>6 hours to 3 days), medium-term (4 to 7 days), and long term prediction (>7days) [11].

In terms of methodology, four groups of RES forecasting approaches can be identified: persistence models, physical models, statistical models, and hybrid models [12].

Most papers on wind power forecasting literature over the last years have focused on different variants of statistical and machine learning approaches, generalized to generate probabilistic forecasts [13]. Therefore, this section will focus mainly on statistical and machine-learning approaches for modeling $f(\cdot)$.

2.2.1 Time Series Models

Time series models are statistical models that use historical data to predict future values of a time series. Several types of time series models may be considered, including the autoregressive model (AR), moving average model (MA), autoregressive moving average model (ARMA), and autoregressive integrated moving average model (ARIMA). The ARMA model is used for wind power forecasting in U.S. wind farms in [14]. The AR model using a Bayesian approach is used to forecast the wind speed in [15].

2.2.2 Machine Learning Models

Linear Regression

The linear regression model to predict the target for agent i ($Y_{i,t}$) using all available data at time t_0 ($\mathcal{Z}_t$) has the form

$$f(\mathcal{Z}_t;\boldsymbol{\beta}^i,\boldsymbol{\eta}^i) = \beta_0 + \underbrace{\overbrace{\sum_{k=1}^{n_i} \beta^i_{k,i} X^i_{j,t}}^{\text{own data}} + \overbrace{\sum_{j=1,j\neq i}^{N} \sum_{k=1}^{n_j} \beta^i_{k,j} X^k_{j,t}}^{\text{others' data}}}_{\text{exogenous variables}} + \underbrace{\overbrace{\sum_{\ell=1}^{L} \eta^i_{\ell,i} Y_{i,t_0-\ell}}^{\text{own lags}} + \overbrace{\sum_{j\neq i}\sum_{\ell=1}^{L} \eta^i_{\ell,j} Y_{j,t_0-\ell}}^{\text{others' lags}}}_{\text{endogenous variables}} + \varepsilon, \tag{2.5}$$

where $\beta^i_{k,j}$ is the coefficient associated to the k-th variable of agent j when predicting $Y_{i,t}$, and $\eta^i_{\ell,j}$ is the coefficient associated to the ℓ most recent power measurement from agent j when predicting $Y_{i,t}$. Both $\boldsymbol{\beta}^i$ and $\boldsymbol{\eta}^i$ are unknown and must be estimated. Consider $\mathbf{z}_t$ and $\mathbf{y}_{i,t}$ are the observed values for $\mathcal{Z}_t$ and $Y_{i,t}$, respectively. The most popular estimation method is *least squares*, in which we pick the coefficients $\boldsymbol{\beta}^i$ and $\boldsymbol{\eta}^i$ that minimize the residual sum of squares

$$\mathrm{RSS}(\boldsymbol{\beta}^i,\boldsymbol{\eta}^i) = \frac{1}{2T}\sum_{t=1}^{T}\left(\boldsymbol{y}_{i,t} - f\left(\mathbf{z}_t;\boldsymbol{\beta}^i,\boldsymbol{\eta}^i\right)\right)^2, \tag{2.6}$$

where $f(\cdot)$ is as defined in (2.5) and T is the number of historical records of the variables. We choose $\boldsymbol{\beta}^i$ and $\boldsymbol{\eta}^i$ that minimize the quantity (2.6), i.e.,

$$\hat{\boldsymbol{\beta}}^{\text{linear}},\hat{\boldsymbol{\eta}}^{\text{linear}} = \underset{\boldsymbol{\beta},\boldsymbol{\eta}}{\operatorname{argmin}}\sum_{t=1}^{T}\left(\boldsymbol{y}_{i,t} - f\left(\mathbf{z}_t;\boldsymbol{\beta},\boldsymbol{\eta}\right)\right)^2 \tag{2.7}$$

This type of model is a simple and straightforward implementation. The coefficients provide a clear understanding of the relationship between predictors and the target variable. For the prediction task of agent i, this method enables us to know which agents

$j, j \neq i$, contributed more to the forecasting. Furthermore, the optimization problem can be solved analytically, making it computationally efficient.

Linear LASSO Regression

Linear LASSO (Least Absolute Shrinkage and Selection Operator) Regression is a shrinkage method that retains a subset of the predictors and discards the rest. In this method, the coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ are estimated in the following way:

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{\text{lasso}}, \hat{\boldsymbol{\eta}}^{\text{lasso}} &= \underset{\boldsymbol{\beta},\boldsymbol{\eta}}{\operatorname{argmin}} \sum_{t=1}^{T} \left(\boldsymbol{y}_{i,t} - f\left(\mathbf{z}_t; \boldsymbol{\beta}, \boldsymbol{\eta}\right) \right)^2 \\ \text{subject to} & \sum_{j=1}^{N} \sum_{k=1}^{n_j} \left| \beta_{k,j}^{i} \right| + \sum_{j=1}^{N} \sum_{\ell=1}^{L} \left| \eta_{\ell,j}^{i} \right| \leq \tau. \end{aligned} \tag{2.8}$$

where τ is the prespecified free parameter that determines the degree of regularization. We can also write the lasso problem in the equivalent *Lagrangian* form:

$$\hat{\boldsymbol{\beta}}^{\text{lasso}}, \hat{\boldsymbol{\eta}}^{\text{lasso}} = \underset{\boldsymbol{\beta},\boldsymbol{\eta}}{\operatorname{argmin}} \left\{ \frac{1}{2T} \sum_{t=1}^{T} \left(\boldsymbol{y}_{i,t} - f\left(\mathbf{z}_t; \boldsymbol{\beta}, \boldsymbol{\eta}\right) \right)^2 + \lambda \left(\sum_{j=1}^{N} \sum_{k=1}^{n_j} \left| \beta_{k,j}^{i} \right| + \sum_{j=1}^{N} \sum_{\ell=1}^{L} \left| \eta_{\ell,j}^{i} \right| \right) \right\}, \tag{2.9}$$

where $f(\cdot)$ is the function in (2.5).

The LASSO penalty in regression introduces sparsity in the estimated coefficients, facilitating automatic feature selection. In machine learning, Bayesian Optimization is a commonly employed technique for hyperparameter tuning [16], which will be discussed in Section 2.2.3. Therefore, Bayesian optimization can be utilized to determine the optimal value of λ. One of the advantages of LASSO Regression is its ability to automatically perform feature selection, enhancing both the interpretability and performance of the model. Additionally, the bias introduced by the regularization penalty can prevent overfitting by shrinking certain coefficients toward zero. This property allows LASSO Regression to effectively exclude irrelevant features, enabling only relevant agents j to contribute to the prediction task of agent i.

Linear and LASSO regression models have their limitations when it comes to capturing nonlinear relationships between variables. In order to address this issue, researchers have explored the use of additive models, which provide a flexible framework for modeling complex relationships. One approach to constructing additive models involves transforming the variables in a manner that enables a linear model in the coefficients to capture the underlying nonlinear relationships. Examples of such transformations include kernel linear regression and spline regression, commonly used in practice.

Splines Regression

Spline regression uses piecewise polynomials to approximate the nonlinear relationship between the predictors $\mathcal{Z}_t$ and the response variable $Y_{i,t}$. In this thesis, each variable $X \in \mathcal{Z}_t$ is transformed into $D+K+1$ new variables, obtained by applying B-spline basis functions of order D with K interior knots. By applying this transformation to all variables in $\mathcal{Z}_t$, we are left with an augmented set of variables $\tilde{\mathcal{Z}}_t$ that can be fed as input into a linear or lasso linear regression model. This transformation captures nonlinear complex relations while providing computational efficiency and numerical stability.

In what follows, a brief description of how to obtain such B-spline basis functions of order D, considering K interior knots, is presented. The approach involves dividing the domain of each explanatory variable $X \in \mathcal{Z}_t$ into several intervals $]-\infty, \tau_0]$, ..., $[\tau_{K+1}, \infty[$, where

$$\tau_0 < \tau_1 < \ldots < \tau_K < \tau_{K+1}$$

are points called knots. $\tau_1, \ldots, \tau_K$ are called interior knots and are typically selected as the quantiles from the empirical distribution of the underlying variable. Consider the augmented knot set given by

$$\tau_{-D} = \cdots = \tau_0 \leq \tau_1 \leq \cdots \leq \tau_{K+1} = \cdots = \tau_{K+D+1}, \tag{2.10}$$

where the lower and upper boundary knots τ_0 and τ_{K+1} are appended D times due to the recursive nature of the B-spline basis functions of order D. Usually, an index reset is applied so that the $K+2(D+1)$ augmented knots in (2.10) are now indexed by $i \in \{0, \ldots, K+2D+1\}$.

For each of the augmented knots $\tau_i, i \in \{0, \ldots, K+2D+1\}$, a set of functions $B_{i,d}(x)$ is recursively defined, $d \in \{0, \ldots, D\}$, where D is the degree of the B-spline basis as follows:

$$B_{i,1}(x) = \begin{cases} 1 & \text{if } \tau_i \leq x < \tau_{i+1} \\ 0 & \text{otherwise,} \end{cases} \tag{2.11}$$

$$B_{i,d}(x) = \frac{x - \tau_i}{\tau_{i+d-1} - \tau_i} B_{i,d-1}(x) + \frac{\tau_{i+d} - x}{\tau_{i+d} - \tau_{i+1}} B_{i+1,d-1}(x). \tag{2.12}$$

Each variable X is therefore transformed into $K + D + 1$ variables given by $[B_{0,D}(x), \ldots, B_{K+D,D}(x)]$.

Figure 2.1 illustrates the B-spline basis functions obtained when considering $K = 2$ interior knots, given by (3.33, 6.46), the boundary knots (0.20, 9.60), and the degree of the B-spline basis is $D = 3$.

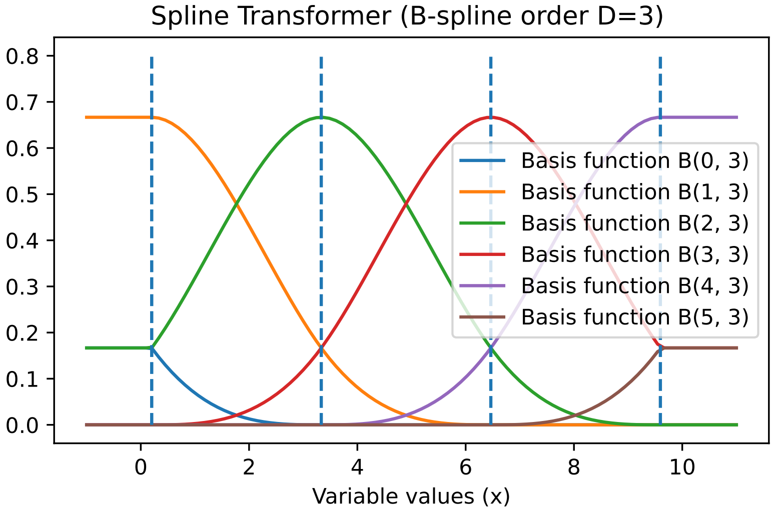


FIGURE 2.1: Example of B-spline basis function of order $D = 3$, and two interior knots ($K = 2$).

In this thesis, the number of knots K, and the B-spline degree D are hyperparameters chosen such that the residual sum of squares is minimized:

$$\text{RSS}(\boldsymbol{\beta}) = \frac{1}{2T} \sum_{t=1}^{T} \left(y_{i,t} - f(\tilde{\mathbf{z}}_{j,t}; \boldsymbol{\beta}) \right)^2, \tag{2.13}$$

where $\tilde{z}_{j,t} \in \tilde{\mathcal{Z}}$ and $\tilde{\mathcal{Z}}$ is the augmented spline data set that can be fed into a Linear Regression (Section 2.2.2) or a LASSO Regression model (Section 2.2.2).

Gradient Boosting Regressor

Gradient Boosting is a popular machine-learning technique for classification and regression tasks. In the context of regression, the Gradient Boosting Regressor (GBR) combines multiple weak learners (typically decision trees) to create a strong predictive model.

Boosting algorithms operate by sequentially training weak learners, with each learner aiming to rectify its predecessor's mistakes. Consequently, the algorithm consistently acquires knowledge that may not be entirely precise but constitutes a gradual advancement in the correct path. As the algorithm progresses by iteratively addressing previous errors, its predictive capability becomes increasingly refined. These weak learners can be, for example, regression trees.

Regression trees divide the input space in disjoint regions $\boldsymbol{R}_j \in \mathbb{R}^D, j \in \{1, 2, \ldots, J\}$. A constant function γ_j is assigned to each such region, and the predictive rule is

$$\boldsymbol{z}_t \in \boldsymbol{R}_j \Rightarrow f(\boldsymbol{z}_t) = \gamma_j. \tag{2.14}$$

Thus, a tree can be formally expressed as

$$\mathcal{T}(\boldsymbol{z}_t; \Theta) = \sum_{j=1}^{J} \gamma_j \mathcal{I}\left(\boldsymbol{z}_t \in \boldsymbol{R}_j\right), \tag{2.15}$$

with parameters $\Theta = \left\{\boldsymbol{R}_j, \gamma_j\right\}_1^J$. The indicator function $\mathcal{I}(\cdot)$ is defined as:

$$\mathcal{I}(\text{condition}) = \begin{cases} 1, & \text{if condition is true} \\ 0, & \text{if condition is false} \end{cases} \tag{2.16}$$

The parameters are found by minimizing the empirical risk [17], which is the sum of the loss function applied to the predictions and the corresponding true target values:

$$\hat{\Theta} = \arg\min_{\Theta} \sum_{j=1}^{J} \sum_{\boldsymbol{z}_t \in \boldsymbol{R}_j} \mathcal{L}\left(y_{i,t}, \gamma_j\right) \tag{2.17}$$

If we choose $\mathcal{L}(y_t, \gamma_j) = (y_t - \gamma_j)^2$ and the regions R_j are defined, the optimal γ_{jm} values are found by differentiating the empirical risk with respect to γ_{jm} and set it to zero:

$$\frac{\partial}{\partial \gamma_j} \sum_{z_t \in R_j} \mathcal{L}(y_t, \gamma_j) = 0 \tag{2.18}$$

which holds the result

$$\gamma_j = \frac{\sum_{z_t \in R_j} y_{i,t}}{n_j} \tag{2.19}$$

i.e. the solution to the minimization of the squared error loss function leads to coefficients γ_{jm} equal to the mean of the residuals in each region. However, finding the best region partitioning is not solved analytically. Instead, a heuristic search procedure has to be used. A typical strategy is to use a greedy, top-down induction of decision trees algorithm.

The gradient boosting model is a sum of such trees

$$f_M(\boldsymbol{z}_t) = \sum_{m=1}^{M} \underbrace{\mathcal{T}\left(\boldsymbol{z}_t; \Theta_m\right)}_{f_m} \tag{2.20}$$

induced in a forward stagewise manner:

$$f_m(z_t) = f_{m-1}(z_t) + \mathcal{T}\left(z_t; \Theta_m\right) \tag{2.21}$$

At each iteration m, one must solve

$$\hat{\Theta}_m = \arg\min_{\Theta_m} \sum_{t=1}^{T} \mathcal{L}\left(y_{i,t}, f_{m-1}\left(z_t\right) + \mathcal{T}\left(x_i; \Theta_m\right)\right) \tag{2.22}$$

for the region set and constants $\Theta_m = \{\boldsymbol{R}_{jm}, \gamma_{jm}\}_1^{J_m}$ of the next tree $\mathcal{T}\left(z_t; \Theta_m\right)$, given the current model $f_{m-1}(z_t)$.

Once the regions R_{jm} are defined, the optimal constants γ_{jm} are computed by minimizing the loss function:

$$\hat{\gamma}_{jm} = \arg\min_{\gamma_{jm}} \sum_{z_t \in R_{jm}} \mathcal{L}\left(y_{i,t}, f_{m-1}\left(z_t\right) + \gamma_{jm}\right). \tag{2.23}$$

However, the minimization problem (2.22) is, in general, an optimization problem computationally infeasible. An approximation is implemented instead, using gradient descent. To find a local minimum of the loss function, the steepest descent step $-\eta_m * r_m$ is applied to the minimization problem, where η_m is the step length and

$$r_m = \nabla f_{m-1}(z_t) \tag{2.24}$$

In fact, the negative gradient is the maximal descent direction. Therefore, at each iteration, the model is updated:

$$f_m = f_{m-1} - \eta_m r_m \tag{2.25}$$

and the process is repeated in the next iteration. However, the gradient (2.24) is only defined in the training data points $\{z_t\}_{t=1}^{T}$, whereas the ultimate goal is to generalize f_M to H timesteps ahead $\{z_t\}_T^{T+H}$. To solve this, a tree $\mathcal{T}(z_t, \Theta)$ is fitted to the negative gradient values (2.24) using the *least squares*:

$$\tilde{\Theta}_m = \arg_{\Theta} \min \sum_{i=1}^{N} \left(-r_m - T\left(z_t; \Theta\right)\right)^2 \tag{2.26}$$

and the regions $\tilde{R}_{jm}$ become determined. One is now able to compute the constants γ_{jm} given by (2.23), which is reduced to the solution (2.19) when we choose the lost function to be $\mathcal{L}(y_{i,t}, \tilde{\Theta}_m) = \frac{1}{T}\left[y_{i,t} - f_m(z_t)\right]^2$.

2.2.3 Bayesian Optimization

Bayesian optimization is a sequential model-based optimization technique that solves expensive and black-box optimization problems, such as hyperparameter tuning in machine learning.

The Bayesian optimization algorithm seeks to discover the optimal parameters $\boldsymbol{\theta}$ that minimize a single-valued objective function $l(\boldsymbol{\theta})$, such as cross-validation mean squared error. There are two primary choices when performing Bayesian optimization: a prior distribution of the objective function and an acquisition function. Generally, Bayesian optimization employs a Gaussian process (GP) as a prior distribution. A brief description of Gaussian processes and Acquisition functions is presented next, followed by the Bayesian optimization algorithm.

Gaussian Process

The Gaussian Process provides a way to model the probability distribution of possible values, $l(\boldsymbol{\theta})$, for a given function $l(\cdot)$ at each point $\boldsymbol{\theta}$. These probability distributions are characterized by Gaussian distributions, where the mean $\mu(\boldsymbol{\theta})$ and variance $\sigma^2(\boldsymbol{\theta})$ may vary for different $\boldsymbol{\theta}$. Consequently, we define a probability distribution over functions as follows:

$$P(l(\boldsymbol{\theta})|\boldsymbol{\theta}) = \mathcal{N}(\mu(\boldsymbol{\theta}), \sigma^2(\boldsymbol{\theta})), \tag{2.27}$$

where $\mathcal{N}$ represents the standard normal distribution.

To estimate the Gaussian process, given a set of S observations $\mathcal{S}_{1:S} = \{l(\boldsymbol{\theta}_i)\}_{i=1}^{S}$ and a user-specified sampling noise σ^2_{noise}, the following steps are taken:

$$P(l(\boldsymbol{\theta})|\mathcal{S}_{1:S}, \boldsymbol{\theta}) = \mathcal{N}\left(\mu_S(\boldsymbol{\theta}), \sigma_S^2(\boldsymbol{\theta})\right), \tag{2.28}$$

where

$$\mu_S(\boldsymbol{\theta}) = \mathbf{k}^\top \mathbf{K}^{-1} \mathcal{S}_{1:S}, \tag{2.29}$$

$$\sigma_S^2(\boldsymbol{\theta}) = k(\mathbf{x}, \mathbf{x}) - \mathbf{k}^\top \mathbf{K}^{-1} \mathbf{k}, \tag{2.30}$$

Here, $\mathbf{K}$ represents the kernel matrix given by

$$\mathbf{K} = \begin{bmatrix} k(\boldsymbol{\theta}_1, \boldsymbol{\theta}_1) & \dots & k(\boldsymbol{\theta}_1, \boldsymbol{\theta}_S) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{\theta}_S, \boldsymbol{\theta}_1) & \dots & k(\boldsymbol{\theta}_S, \boldsymbol{\theta}_S) \end{bmatrix} + \sigma^2_{\text{noise}} \mathbf{I}, \tag{2.31}$$

and $\mathbf{k} = [k(\boldsymbol{\theta}, \boldsymbol{\theta}_1), k(\boldsymbol{\theta}, \boldsymbol{\theta}_2), \cdots, k(\boldsymbol{\theta}, \boldsymbol{\theta}_S)]$. A commonly used option for k is the radial basis function kernel (RBF) given by $k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma}\right)$.

Acquisition Function

In the search for the next point $\boldsymbol{\theta}_{S+1}$ to evaluate, various acquisition functions can be employed, including the probability of improvement, the expected improvement, or the Lower Confidence Bound. As an example, the equation for the Lower Confidence Bound is as follows:

$$\boldsymbol{\theta}_{S+1} = \arg\max_{\boldsymbol{\theta}} \left(2\sigma_S(\boldsymbol{\theta}) - \mu_S(\boldsymbol{\theta})\right). \tag{2.32}$$

Bayesian optimization algorithm

The Bayesian optimization algorithm using Gaussian Processes for hyperparameter tuning is summarized in Algorithm 1. This was the mechanism employed for hyperparameter optimization in the forecasting models presented previously.

Algorithm 1 Bayesian optimization algorithm for hyperparameter tuning.

```
Input: $S$, $\boldsymbol{\theta}_{\min}$, $\boldsymbol{\theta}_{\max}$
Generate randomly $\{\boldsymbol{\theta}_i\}_{i=1}^{S} \in [\boldsymbol{\theta}_{\min}, \boldsymbol{\theta}_{\max}]$
Compute $\mathcal{S}_{1:S} = \{L(\boldsymbol{\theta}_i)\}_{i=1}^{S}$ and estimate $\mu_S(\boldsymbol{\theta})$ and $\sigma_S(\boldsymbol{\theta})$ through (2.29) and (2.30)
while stop criteria not achieved (e.g., number of iterations) do
    Find the new hyperparameter configuration $\boldsymbol{\theta}_{S+1}$ that maximizes the acquisition function, as in (2.32).
    Evaluate the performance of the model with the hyperparameter configuration $\boldsymbol{\theta}_{S+1}$ by training and cross-validating the model.
    Incorporate the new observation $\boldsymbol{\theta}_{S+1}$ and its corresponding performance value $l(\boldsymbol{\theta}_{S+1})$ into the set of observations $\mathcal{S}_{1:S+1}$.
    Update the Gaussian process model of $l(\boldsymbol{\theta})$, i.e., $\mu_{S+1}(\boldsymbol{\theta})$ and $\sigma_{S+1}(\boldsymbol{\theta})$
end while
```

2.2.4 Feature Selection

In the context of supervised learning, feature selection consists of selecting a subset of the n original features which contribute the most to the prediction of the variable we are interested in. In general, by reducing the dimensionality of our regression task, we can reduce the computational cost of training a model and its complexity. Feature selection can also improve the predictor performance, simplify data visualization and facilitate interpretability [18].

The existing approaches are commonly divided into three groups [19]:

1. **Wrapper methods:** use a forecasting model to select feature subsets. Based on the inference from the previous model, wrapper methods decide to add or remove features to the features subset. Some examples of wrapper methods include forward feature selection, backward feature elimination, and recursive feature elimination [20].

2. **Filter methods:** use a similarity metric between the explanatory and the target feature to remove irrelevant features, being independent of the learning algorithm. Common measures include Pearson correlation (that measures linearity between two variables), Spearman correlation (that measures the monotonicity between two variables), mutual information [21], etc.

3. **Embedded methods:** perform feature selection as part of the model construction process. Common methods include L_1 and L_2 regularization on models' coefficients. The LASSO linear regression mentioned in Section 2.2.2 performs variable selection using a L_1 regularization.

2.3 Data Markets

Data markets serve as platforms facilitating the exchange, acquisition, and sharing of data among multiple stakeholders, ensuring their collective benefit. These markets typically comprise sellers, buyers, and a central market operator responsible for data storage, security, and overall market functioning. Due to the exclusive access granted to the market operator, concerns related to data confidentiality are effectively addressed and mitigated.

Several proposals have been put forth in the literature under this framework. In the subsequent subsections, we will examine the main algorithmic solutions that focus on regression problems.

2.3.1 Cooperative Zero-Regret Auction Mechanism

The data market mechanism proposed by Gonçalves et al. [10] introduces an algorithmic solution for data markets in collaborative forecasting, with a focus on renewable energy forecasting by combining spatio-temporal data. Three main types of agents are considered: data buyers, data sellers, and the data market operator. The data market algorithm is illustrated in Figure 2.2 and operates as follows. At a specific time T, a new session begins, where RES agents submit their historical data $\left\{\mathcal{X}_{j,t}\right\}_{t=1}^{T}$ to the market operator. Some agent i, aims to forecast power measurements for the subsequent H time steps $\{Y_{i,t}\}_{t=T+1}^{T+H}$. The market starts by computing a market price $p_i \in \mathbb{R}^+$ representing the price per unit increase in forecasting accuracy for this session – the computation of p_i is iterative and based on previous price data.

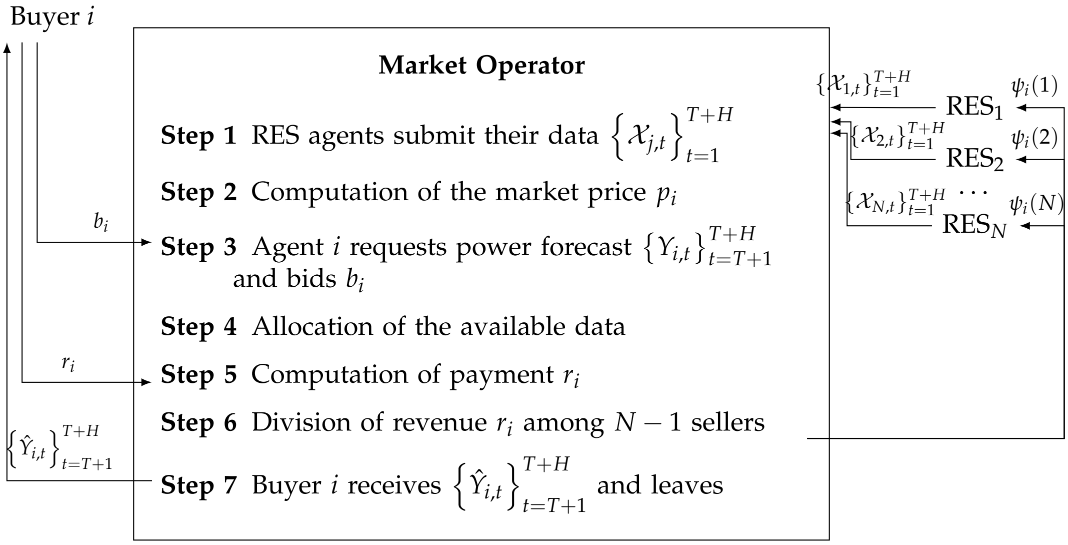

Buyer i
Market Operator
Step 1 RES agents submit their data $\left\{\mathcal{X}_{j,t}\right\}_{t=1}^{T+H}$
Step 2 Computation of the market price p_i
b_i
Step 3 Agent i requests power forecast $\{Y_{i,t}\}_{t=T+1}^{T+H}$ and bids b_i
Step 4 Allocation of the available data
r_i
Step 5 Computation of payment r_i
Step 6 Division of revenue r_i among $N-1$ sellers
$\left\{\hat{Y}_{i,t}\right\}_{t=T+1}^{T+H}$
Step 7 Buyer i receives $\left\{\hat{Y}_{i,t}\right\}_{t=T+1}^{T+H}$ and leaves
$\{\mathcal{X}_{1,t}\}_{t=1}^{T+H}$
RES$_1$
$\psi_i(1)$
$\{\mathcal{X}_{2,t}\}_{t=1}^{T+H}$
RES$_2$
$\psi_i(2)$
$\{\mathcal{X}_{N,t}\}_{t=1}^{T+H}$
$\cdots$
RES$_N$
$\psi_i(N)$

FIGURE 2.2: Zero-regret data market mechanism at time $t = T$.

Upon arrival, agent i requests a power forecast for the next H time steps and submits a bid (b_i) indicating the amount they are willing to pay per unit of improvement in forecasting accuracy. Considering the bid (b_i) and market price (p_i), the market operator allocates the available features – this allocation process determines how much noise will be introduced to the data provided by the sellers, in case $b_i < p_i$.

Subsequently, the market operator applies cross-validation to estimate the gain in forecasting accuracy when using the collaborative forecast, instead of a local forecast. The final payment (r_i) for agent i is computed accordingly, and distributed among the sellers. The percentage $\psi_i(j)$ for each seller $j \in \mathcal{A} \setminus \{i\}$ relates to the importance of their data. The ideal procedure for assessing the relevance of each feature involves exhaustive training of the statistical model across all possible combinations of features. This method is known as Shapley Allocation [22]. However, to overcome computational complexities, a Shapley Approximation is used where only a subset of all possible features combinations are used. The selection of feature combinations is inherently stochastic, leading to differing subsets being chosen across iterations. Finally, the buyer i receives $\left\{\hat{Y}_{i,t}\right\}_{t=T+1}^{T+H}$ and leaves the market session.

Throughout the process, the sellers continuously update their data as new time steps occur, ensuring that the available information remains relevant and up-to-date.

Noteworthy characteristics of this framework are:

i) Buyers acquire forecasts rather than specific datasets without knowledge of the datasets used in generating the forecasts.

ii) Equitable revenue distribution among sellers with similar information.

iii) A market price determined by the buyer's benefit, ensuring payment only when forecasting skills are improved.

iv) Compensation based on incremental gain.

v) The market operator is model-free, meaning that the buyers can provide their own model, and the market operator can forecast using any forecasting model.

2.3.2 Online Regression Market

The proposal in [23] offers a data framework for data markets in collaborative forecasting, focusing on regression tasks in the energy sector. The authors consider two main components: central and support agents. Figure 2.3 illustrates this proposal. A central agent i takes the buyer role and posts a regression task on the platform. The regression task is requested by i to the market operator and can be performed either as a batch learning task or an online learning task by a regression model $f_i(\cdot)$ specified by i. The central agent $i \in \mathcal{A}$ also expresses its willingness to pay, denoted as b_i, for improving the model fitting or forecast accuracy, as measured by a loss function $\ell(\cdot)$. The payment for the i-th agent can be expressed as follows:

$$p_i = \underbrace{(\mathcal{L}_{\text{local}} - \mathcal{L}_{\text{market}})}_{\text{loss improvement with market data}} \times b_i, \tag{2.33}$$

where the loss values, $\mathcal{L}_{\text{local}}$ and $\mathcal{L}_{\text{market}}$, are obtained when predicting $Y_{i,t}$ using $X_{i,t}$ and $\mathcal{Z}_t$, respectively. The value of b_i is stated in monetary terms per unit improvement in the loss function $\ell(\cdot)$ and per data point provided.

Support agents $j \in \mathcal{A}$, $j \neq i$, are sellers in this data market. They have sets of features $\mathcal{X}_{j,t}$, $j \neq i$, which they share with the analytics platform. At each time t, support agents provide new data to the platform as time progresses. The platform incorporates these

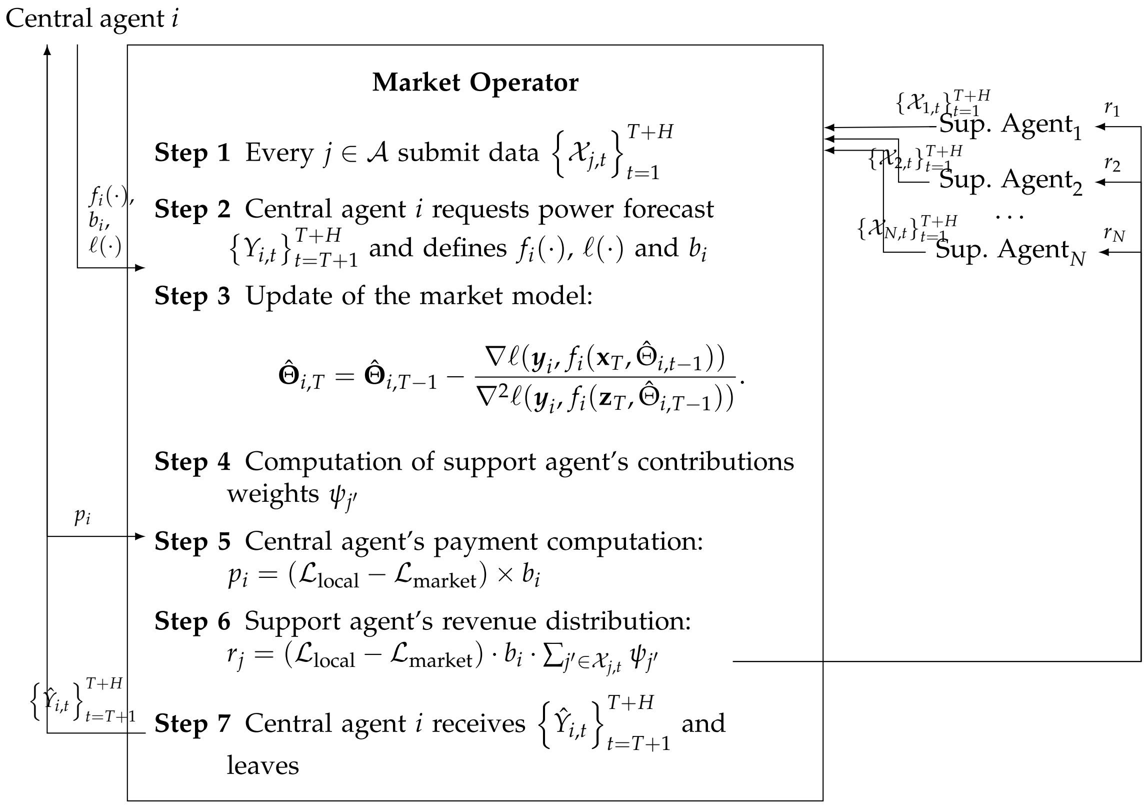

Central agent i
Market Operator
Step 1 Every $j \in \mathcal{A}$ submit data $\left\{\mathcal{X}_{j,t}\right\}_{t=1}^{T+H}$
Step 2 Central agent i requests power forecast $\{Y_{i,t}\}_{t=T+1}^{T+H}$ and defines $f_i(\cdot)$, $\ell(\cdot)$ and b_i
Step 3 Update of the market model:
$\hat{\mathbf{\Theta}}_{i,T} = \hat{\mathbf{\Theta}}_{i,T-1} - \frac{\nabla \ell(\boldsymbol{y}_i, f_i(\mathbf{x}_T, \hat{\mathbf{\Theta}}_{i,t-1}))}{\nabla^2 \ell(\boldsymbol{y}_i, f_i(\mathbf{z}_T, \hat{\mathbf{\Theta}}_{i,T-1}))}.$
Step 4 Computation of support agent's contributions weights $\psi_{j'}$
Step 5 Central agent's payment computation: $p_i = (\mathcal{L}_{\text{local}} - \mathcal{L}_{\text{market}}) \times b_i$
Step 6 Support agent's revenue distribution: $r_j = (\mathcal{L}_{\text{local}} - \mathcal{L}_{\text{market}}) \cdot b_i \cdot \sum_{j' \in \mathcal{X}_{j,t}} \psi_{j'}$
Step 7 Central agent i receives $\left\{\hat{Y}_{i,t}\right\}_{t=T+1}^{T+H}$ and leaves
$f_i(\cdot)$, b_i, $\ell(\cdot)$
p_i
$\left\{\hat{Y}_{i,t}\right\}_{t=T+1}^{T+H}$
$\{\mathcal{X}_{1,t}\}_{t=1}^{T+H}$
Sup. Agent$_1$
r_1
$\{\mathcal{X}_{2,t}\}_{t=1}^{T+H}$
Sup. Agent$_2$
r_2
...
$\{\mathcal{X}_{N,t}\}_{t=1}^{T+H}$
Sup. Agent$_N$
r_N

FIGURE 2.3: Online Regression Market mechanism at time $t = T$.

data from the support agents to continuously learn and update the regression model, and consequently $\mathcal{L}_{\text{local}}$ and $\mathcal{L}_{\text{market}}$.

The platform facilitates the market by matching the regression tasks posted by the central agent with the feature data provided by the support agents. The key goal is to improve the value of the loss function $\ell(\cdot)$ through model updates and incorporating relevant feature data. This improvement in the loss function can lead to remunerations for the support agents based on the contribution of their features to improve the loss function $\ell(\cdot)$. The revenue for the j-th support agent is determined as follows:

$$r_j = \underbrace{(\mathcal{L}_{\text{local}} - \mathcal{L}_{\text{market}})}_{\text{loss improvement with market data}} \times b_i \times \underbrace{\sum_{j' \in \mathcal{X}_{j,t}} \psi_{j'}}_{\text{contribution of } j\text{th agent features}}, \tag{2.34}$$

where $\psi_{j'}$ measures the contribution of the j'-th variable from the j-th support agent to improve the loss function $\ell(\cdot)$, $\psi_{j'} \in [0,1]$ and $\sum_j \sum_{j' \in \mathcal{X}_{j,t}} \psi_{j'} = 1$. Shapley values are considered to determine such weights $\psi_{j'}$ – see [23] for more details.

The framework also includes an online version of the regression market to adapt to the streaming nature of data and the need for continuous learning in an online environment. In this online setting, the central agent still expresses their willingness to pay b_i for reducing the value of the loss function $\ell(\cdot)$. The central agent specifies the regression model f_i, their own set of features $\mathcal{X}_{i,t}$, and the duration over which the learning process will occur.

The support agents continue to provide their sets of feature data at each time instant, and the platform defines a mapping $f_i\left(\mathcal{Z}_t, \Theta_i\right)$ within the analytics platform to accommodate the online regression market. The online market allows for recursive updates of the regression model parameters based on newly collected data, using a Newton-Raphson step:

$$\hat{\boldsymbol{\Theta}}_{i,t} = \hat{\boldsymbol{\Theta}}_{i,t-1} - \frac{\nabla \ell(\boldsymbol{y}_{i,t}, f_i(\mathbf{x}_t, \hat{\Theta}_{i,t-1}))}{\nabla^2 \ell(\boldsymbol{y}_{i,t}, f_i(\mathbf{x}_t, \hat{\Theta}_{i,t-1}))}. \tag{2.35}$$

Note that the assumption of twice differentiability of the loss function is made. For example, for L^2 norm, we have $\ell(\boldsymbol{y}_{i,t}, f_i(\mathbf{x}_t, \Theta_{i,t}))$ enables this updating method, which has been proven to be feasible and experimentally confirmed.

Moreover, the online configuration incorporates a fading window using a forgetting factor (λ) to assign more importance to recent data. This ensures the model adapts to changing patterns and dynamics in the energy sector by assigning a higher weight to more recent observations.

2.3.3 LASSO Regression Market

The framework proposed in [24] explores the data market design of [25] by introducing a regression market for wind agents to trade wind power data that allows the sellers to customize their data payments within predefined limits.

In the proposed data market mechanism, there is a central agent denoted by $i \in \mathcal{A}$, that acts as a buyer. Figure 2.4 illustrates this proposal. The central agent's task is to predict its target variable $\left\{Y_{i,t}\right\}_{t=T+1}^{T+H}$ by using the available data in the market $\mathcal{Z}_t$. The other agents in the market are called support agents, the same as sellers in the previous frameworks.

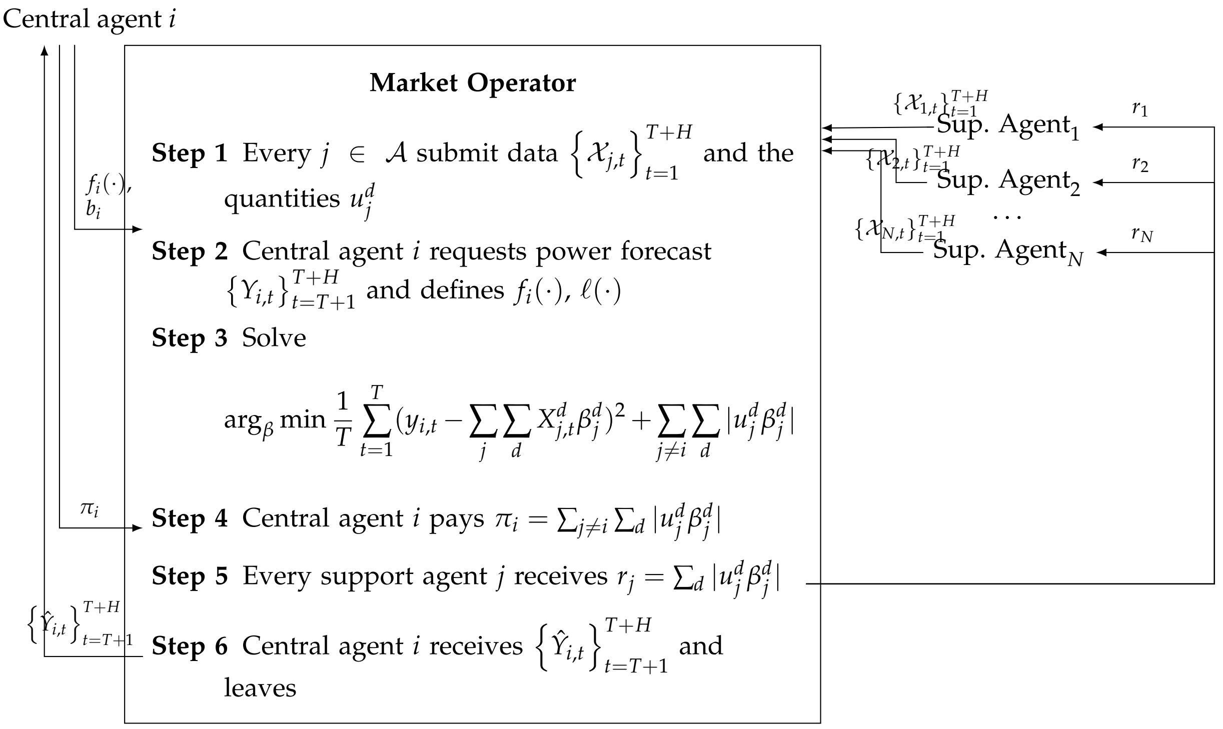
It is introduced a payment threshold, denoted as $H_j^d(u_j^d, \beta_j^d)$, which represents the payment required by a support agent $j \in \mathcal{A}$, $j \neq i$ for disclosing data associated with their dth feature. The coefficient β_j^d measures the relation of agent j's dth feature with the central agent's target variable y_i. The seller determines the quantity u_j^d that reflects its reservation to sell the specific feature, taking into account factors such as the potential loss of accuracy, loss due to an increase in competitors' profit, data collection costs, and other considerations.

The forecasting framework is constructed based on a lasso regression model, as the one explained in Section 2.2.2,

$$\arg_\beta \min \frac{1}{T}\sum_{t=1}^{T}(y_{i,t} - \sum_j \sum_d X_{j,t}^d \beta_j^d)^2 + \overbrace{\sum_{j\neq i} \underbrace{\sum_d |u_j^d \beta_j^d|}_{\text{value received by } j\text{th seller}}}^{\text{value paid by } i\text{th buyer}} . \tag{2.36}$$

Each seller can set the lasso parameter λ_j^d based on their reservation value u_j^d. This allows agents to balance their willingness to share data with the impact on their own models.

In summary, the data market operates as follows: In the initial time step, all agents submit their data, and support agents indicate their willingness to share data, denoted as u_i^d, to the central agent i. The market operator computes the minimization lasso problem in (2.36), considering the lasso terms and the reservations of support agents. The final payment is made to each support agent based on the product of their reservation to sell a specific feature and the absolute value of the corresponding estimated coefficient. This market mechanism has been proven to meet the profit requirements of support agents while ensuring financial benefits for the central agent.


Central agent i
$f_i(\cdot)$, b_i
π_i
$\left\{\hat{Y}_{i,t}\right\}_{t=T+1}^{T+H}$
$\{\mathcal{X}_{1,t}\}_{t=1}^{T+H}$
$\{\mathcal{X}_{2,t}\}_{t=1}^{T+H}$
$\{\mathcal{X}_{N,t}\}_{t=1}^{T+H}$
Sup. Agent$_1$
Sup. Agent$_2$
...
Sup. Agent$_N$
r_1
r_2
r_N

Market Operator

Step 1 Every $j \in \mathcal{A}$ submit data $\left\{\mathcal{X}_{j,t}\right\}_{t=1}^{T+H}$ and the quantities u_j^d

Step 2 Central agent i requests power forecast $\left\{Y_{i,t}\right\}_{t=T+1}^{T+H}$ and defines $f_i(\cdot)$, $\ell(\cdot)$

Step 3 Solve

$$\arg_\beta \min \frac{1}{T}\sum_{t=1}^{T}(y_{i,t} - \sum_j \sum_d X_{j,t}^d \beta_j^d)^2 + \sum_{j\neq i}\sum_d |u_j^d \beta_j^d|$$

Step 4 Central agent i pays $\pi_i = \sum_{j\neq i}\sum_d |u_j^d \beta_j^d|$

Step 5 Every support agent j receives $r_j = \sum_d |u_j^d \beta_j^d|$

Step 6 Central agent i receives $\left\{\hat{Y}_{i,t}\right\}_{t=T+1}^{T+H}$ and leaves

FIGURE 2.4: LASSO Regression Market mechanism at time $t = T$.

Chapter 3

Proposal

Our proposed framework introduces a data market tailored for collaborative forecasting within the energy sector. This innovative marketplace is characterized by its bid-based interaction approach, strategically designed to incentivize active participation from both buyers and sellers. The ultimate goal of our market is to enhance forecasting accuracy through collective intelligence. Although this data market can be implemented in other sectors, we present our proposal adapted to RES Forecasting while employing the formulation described in Section 2.1. The rest of this chapter is structured as follows. Section 3.1 introduces the key entities in this market, Section 3.2 outlines the bidding process and price formation, Section 3.3 details the modeling of linear spline regression under budget constraints, and Section 3.4 synthesizes these steps to elucidate the proposed auction mechanism.

3.1 Data Market Entities

Like established data markets in Section 2.3, our framework comprises three fundamental entities: buyers, sellers, and the market operator.

A seller is an agent $j \in \mathcal{S} \subseteq \mathcal{A}$ aiming to generate monetary revenue by submitting their data sequences $\mathcal{X}_{j,t}$ to the market. These sellers are motivated by revenue generation and lack specific knowledge about the forecasting methods that will use their data.

On the other side, a buyer is an agent $i \in \mathcal{B} \subseteq \mathcal{A}$ with a unique regression task $\{Y_{i,t'}\}_{t'=t+1}^{t+H}$. Without the collaborative framework, a buyer i would be constrained to approximate $Y_{i,t}$ using a local function $f^{(i)}_{\text{local}}(\boldsymbol{X}_i, \boldsymbol{\Theta}_i)$ built on local historical data $\{\mathcal{X}_{i,t}\}_{t=1}^{T}$. Here, $\boldsymbol{X}_i$ is the correspondent local feature matrix and $\boldsymbol{\Theta}_i$ represents the parameters to be estimated. However, with our market model, buyer i enters the market intending to enhance forecasting accuracy through the use of the market model $f^{(i)}_{\text{market}}(\mathbf{Z}, \mathbf{\Theta}_{\boldsymbol{i}})$, where $\mathbf{Z}$ is the feature matrix that incorporates all historical data collected by the market $\{\mathcal{Z}_t\}_{t=1}^{T}$.

At the center of this dynamic interaction lies the market operator, responsible for orchestrating a range of critical operations. This includes acquiring data from sellers, managing an auction mechanism incorporating both buyers and sellers, performing regression tasks, computation of payment prices, equitable revenue distribution among sellers, and overseeing other tasks to ensure the smooth functioning of the market. Sellers $j \in \mathcal{S}$ only possess access to their individual data $\mathcal{X}_{j,t}$, while buyers $i \in \mathcal{B}$ can only access power forecasts specifically generated for their respective power plants, $Y_{i,t}$. This setup ensures data privacy, contingent upon the market operator's integrity and impartiality as a trusted intermediary.

One very important aspect of our framework is that it has the restriction of $f^{(i)}_{\text{market}}$ always being a model linear in its parameters. Conversely, $f^{(i)}_{\text{local}}$ can encompass a wide range of statistical models, chosen at the discretion of buyer i.

Gain Function $\mathcal{G}$

A pivotal component of our framework is the Gain Function $\mathcal{G}_i$, designed to quantify the incremental improvement in forecasting skill achieved by using market data. This gain is typically measured using a loss function $\mathcal{L}(\cdot)$. Within our framework, the loss function needs to be a convex twice-differentiable function with continuous gradient that satisfies the following Lipschitz condition:

$$\left| \frac{\partial \mathcal{L}}{\partial f}\left(y, f_1\right) - \frac{\partial \mathcal{L}}{\partial f}\left(y, f_2\right) \right| \leq K_1 \left| f_1 - f_2 \right|, \tag{3.1}$$

for any y, f_1, f_2 and $K_1 \in \mathbb{R}^+$ and

$$\frac{\partial^2 \mathcal{L}(y, f)}{\partial f^2} \leq K_2, \tag{3.2}$$

for $K_2 \in \mathbb{R}^+$. This is imperative to the construction of our forecasting mechanism that incorporates bid restrictions. One example of such loss function is the quadratic least squares (2.6) whose proof is discriminated in A.

When a buyer $i \in \mathcal{B}$ enter the market at a timestamp $t = T$, the actual gain that a buyer can achieve in forecasting $\{Y_{i,t'}\}_{t'=t+1}^{t+H}$ is not directly attainable due to the uncertainty of future outcomes. To tackle this, we divide the historical data into a training dataset and a validation dataset. Specifically, the explanatory validation set for the local model is denoted as $\boldsymbol{X}_{\boldsymbol{i}}^{(\text{val})}$, containing the last Δ observations of $\boldsymbol{X}_{\boldsymbol{i}}$. The explanatory validation set for the market model is denoted as $\boldsymbol{Z}^{(\text{val})}$, containing the last Δ observations of $\boldsymbol{Z}$. The target validation set is denoted as $\boldsymbol{y}_{\boldsymbol{i}}^{(\text{val})}$, containing the last Δ observations of $\boldsymbol{y}_{\boldsymbol{i}}$. Similarly, the explanatory training set for the local model is $\boldsymbol{X}_{\boldsymbol{i}}^{(\text{tr})}$, containing the remaining $T - \Delta$ observations of $\boldsymbol{X}_{\boldsymbol{i}}$. The explanatory training set for the market model is $\boldsymbol{Z}^{(\text{tr})}$, containing the remaining $T - \Delta$ observations of $\boldsymbol{Z}$.

With these components, the gain in forecasting task of agent i at time $t = T+1, \ldots, T+H$ is measured by its marginal profit quantified by the Gain Function $\mathcal{G}_i$ defined as:

$$\mathcal{G}(\boldsymbol{X}_i, \boldsymbol{Z}, \boldsymbol{y}_i, f_{\text{local}}^{(i)}, f_{\text{market}}^{(i)}) = \frac{\max\left(\mathcal{L}(\boldsymbol{X}_i, \boldsymbol{y}_i, f_{\text{local}}^{(i)}) - \mathcal{L}(\boldsymbol{Z}, \boldsymbol{y}_i, f_{\text{market}}^{(i)}), 0\right)}{\mathcal{L}\left(\boldsymbol{X}_i, \boldsymbol{y}_i, f_{\text{local}}^{(i)}\right)} \times 100 \tag{3.3}$$

where the training-validation division is implicit. This formulation assumes that the buyer's forecasting skill cannot decrease upon entering the market.

3.2 Bids and price definition

Our data market introduces a distinctive bid-based interaction framework to foster buyer and seller engagement within the marketplace. We will dive into how the data market entities are involved in this auction based mechanism.

3.2.1 Buyers

On the one hand, we are interested in ensuring the buyers interests are met. The framework described in Section 2.3.1 employs an idea that allows each buyer $i \in \mathcal{B}$, to propose a public bid, b_i, representing the maximum price they are willing to pay for a unit increase in the gain function $\mathcal{G}(\cdot)$. It signifies their willingness to pay a maximum amount of $\mathcal{G}(\boldsymbol{X}_i, \boldsymbol{Z}, \boldsymbol{y}_i, f_{\text{local}}^{(i)}, f_{\text{market}}^{(i)}, b_i) \times b_i$. However, a buyer valuation of gain may not be so straightforward. For instance, a buyer may assign higher values to the initial gains, signaling their eagerness to pay more for substantial improvements in their forecasting accuracy. As the gain increases, the buyer's valuation of further improvements might diminish, reflecting a diminishing marginal utility of gain. Our framework introduces a more flexible and nuanced way to buyers express their preferences and willingness to pay for improvements in forecasting skill. Instead of providing a fixed set of potential bids, each buyer now offers a unique value function $\mathcal{VF}_i(g)$ that quantifies their bid as a function of the gain achieved through the market data.

3.2.2 Sellers

Sellers, represented by $j \in \mathcal{S}$, are afforded the ability to propose sets of bids $\boldsymbol{s}_j = (s_{j,1}, s_{j,2}, ..., s_{j,n_j})$ for the use of each of their variables. Each $s_{j,k}$ denotes the minimum monetary compensation sellers require for utilizing their k^{th} variable within $\mathcal{X}_{j,t}$,. This means that the condition to use the k^{th} variable of seller j is $r_{j,k} \geq s_{j,k}$, where $r_{j,k}$ is the monetary compensation agent j gets for the use of its k^{th} variable.

3.2.3 Market Operator

The Market Operator plays a pivotal role in managing the auction mechanism and determining the prices that buyers (denoted as p_i for buyer i) should pay. To bridge the gap between what the market offers and the value perceived by the buyers, the Market Operator constructs a Bid-Gain Option Table ($BGOT$) 3.1. This table relates a set of potential bids, denoted as $\tilde{\boldsymbol{b}}$, to their corresponding gains. In essence, it calculates the gain associated with each bid b within the set $\tilde{\boldsymbol{b}}$. This $BGOT$ is a critical component in determining the final bid, p_i, for each individual buyer.

TABLE 3.1: Bid-Gain Option Table ($BGOT$)

Bid ($\tilde{b}$)	Cross-Validated Gain
b_{min}	$\mathcal{G}(\boldsymbol{X}_i, \boldsymbol{Z}, \boldsymbol{y}_i, f^{(i)}_{\text{local}}, f^{(i)}_{\text{market}}, b_{min})$
$b_{min} + \delta_b$	$\mathcal{G}(\boldsymbol{X}_i, \boldsymbol{Z}, \boldsymbol{y}_i, f^{(i)}_{\text{local}}, f^{(i)}_{\text{market}}, b_{min} + \delta_b)$
$b_{min} + 2\delta_b$	$\mathcal{G}(\boldsymbol{X}_i, \boldsymbol{Z}, \boldsymbol{y}_i, f^{(i)}_{\text{local}}, f^{(i)}_{\text{market}}, b_{min} + 2\delta_b)$
...	...
b_{max}	$\mathcal{G}(\boldsymbol{X}_i, \boldsymbol{Z}, \boldsymbol{y}_i, f^{(i)}_{\text{local}}, f^{(i)}_{\text{market}}, b_{max})$

In this bidding process, the revenue that buyer i is willing to pay must cover the bids from sellers, represented as s_{jk}. The minimum value that b_i could assume is b_{min}, which corresponds to the scenario where the Market Operator solely uses the k^{th} variable of agent j (where $j \neq i$) and the local variables $\mathcal{X}_{i,t}$. Conversely, the maximum value for b_i is $b_{max} = \sum_{j=1}^{N} \sum_{k=1}^{n_j} s_{j,k}$, representing the case where the Market Operator utilizes all available market data $\mathcal{Z}_t$. To define the set of potential bids, we introduce δ_b as the minimum non-zero sum between any pair of seller bids ($\min\{s_{jk} + s_{j'k'} : s_{jk} \neq s_{j'k'}\}$). Thus, we define the set $\tilde{\boldsymbol{b}}$ as follows:

$$\tilde{\boldsymbol{b}} = [b_{min}, b_{min} + \delta_b, b_{min} + 2\delta_b, \ldots, b_{max}] \tag{3.4}$$

The final bid b_i for each buyer is determined by finding the point in $BGOT$ with the highest gain value such that their Valuation Function ($\mathcal{VF}(\cdot)$) bid is lower or equal to the bid of the $BGOT$ point. In other words, the final price will be the price that maximizes the gain function $\mathcal{G}(\boldsymbol{\Psi}_i, p)$, where $\boldsymbol{\Psi}_i = \left(\boldsymbol{X}_i, \boldsymbol{Z}, \boldsymbol{y}_i, f^{(i)}_{\text{local}}, f^{(i)}_{\text{market}}\right)$, and that is covered by the value function $\mathcal{VF}_i(\mathcal{G}(\boldsymbol{\Psi}_i, p))$. Therefore, the market price to be payed by buyer i is given by:

$$p_i = \arg_b \max \mathcal{G}(\boldsymbol{\Psi}_i, b) \quad \text{such that} \quad b \leq \mathcal{VF}(\mathcal{G}(\boldsymbol{\Psi}_i, b)). \tag{3.5}$$

The point in question corresponds to the optimal bid that maximizes gains, as determined by the buyer's valuation function. This concept is visually depicted in Figure 3.1. The feasible region of the solutions of (3.5) is depicted as yellow as the buyer's acceptance region. In the first plot, buyer i initially overvalues the potential gains attainable through the market. However, as we move further along, the points of $BGOT$ converge to a region of a fixed gain, while the value function, where higher bids will not add more value to the forecasting task. The ultimate price decision is the point within the buyer's acceptance region that yields the highest gain.

In the second plot, we have a scenario where the value of the gains of a buyer in relation to the market bids is not always the same. In this case, the final price has a higher value than in the first place meaning that more features will be used or features with higher value in terms of forecasting. In this case, the convergence point happens to be very close to b_{max} which corresponds to the point where all available features are used. Conversely, in the last plot, buyer i consistently undervalues its gain for every possible bid. In such a situation, buyer i opts to exit the market and refrains from utilizing it for its regression task.

Note that the Market Operator is the ultimate entity to decide the final payment while regarding the private valuation of gain of the buyer. This prevents engagement in unfair strategic practices on the buyers part. Moreover, the market price is intricately linked to the buyer's benefit, meaning the buyer only pays if there is a demonstrable improvement in forecasting skill. Buyers are charged based on incremental gain and for purchasing

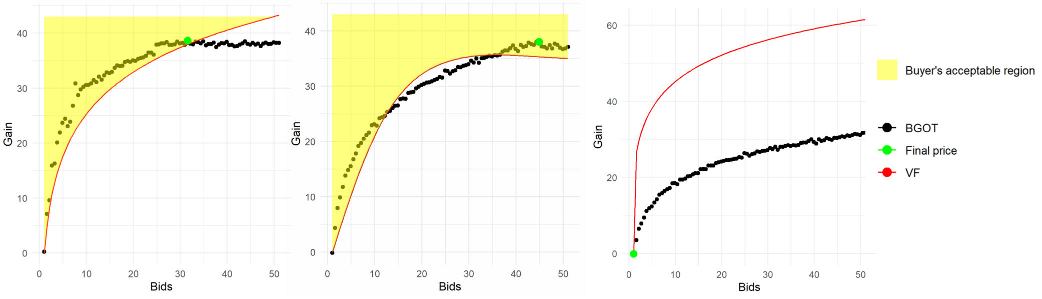


FIGURE 3.1: Illustration of price definition.

forecasts rather than raw features. Importantly, they are not privy to information about which datasets were used to generate these forecasts.

This bid-based approach effectively aligns the incentives of both buyers and sellers, ensuring that both parties strive to extract optimal value from the data exchange.

3.3 Spline Bid-Constrained Lasso Regression

The forecasting problem tackled within our data market needs to account for the restrictions imposed by the buyers and sellers bids and the complex relationships between the data. Generally, when dealing with very short-term predictions, classical time series models (Section 2.2.1) or simpler machine learning models like in Section 2.2.2 can deliver good results in real time. However, when expanding the forecasting time horizon, non linear dependencies need to be captured. A Spline Regressor described in Section 2.2.2 is shown in our case-study (Section 4) to hold very good results while being computationally efficient. Our proposal employs a Spline Lasso Regressor aligned with the auction mechanism described in the previous Sub-section 3.2. We call this model the Spline Bid-Constrained Lasso Regression ($\mathcal{SBCLR}$).

Each variable $z \in \mathcal{Z}_t$ will be transformed into a group of M_z variables, where $M_z = D + K + 1$ is the sum of the spline order D with the number of knots K plus 1 (see Section 2.2.2). After the spline transformer, the original set:

$$\mathcal{Z}_t = \left\{ \underbrace{X^1_{1,t}, \dots, X^{n_1}_{1,t}, Y_{1,t_0-1}, \dots, Y_{1,t_0-L}}_{\text{power plant 1}}, \dots, \underbrace{X^1_{N,t}, \dots, X^{n_N}_{N,t}, Y_{N,t_0-1}, \dots, Y_{N,t_0-L}}_{\text{power plant N}} \right\}. \tag{3.6}$$

is transformed into

$$\tilde{\mathcal{Z}}_t = \left\{ \underbrace{X^1_{1,1,t}, \dots, X^1_{1,M,t}}_{\text{group related to } X^1_{1,t}}, \dots, \underbrace{X^{n_1}_{1,1,t}, \dots, X^{n_1}_{1,M,t}}_{\text{group related to } X^{n_1}_{1,t}}, \dots \right\} \tag{3.7}$$

that is constituted by p groups of variables, where $p = \sum_{j=1}^{N} n_j + NL$ is the number of variables in $\mathcal{Z}_t$.

In this context, we make the assumption that all groups in $\tilde{\mathcal{Z}}_t$ consist of an equal number of variables. This assumption implies that the splines' order D and the number of knots K are equal across all groups of variables. Although in real-world scenarios, this uniformity often doesn't hold true, we adopt this assumption for the sake of simplification.

We need to align the forecasting mechanism with the bids of the participants in the market. We consider seller j wants to receive $s_{j,k}$ if its k^{th} variable is used. Since we apply splines, we consider that if at least one of the variables in the group related to the k^{th} variable is used, then seller j will receive $s_{j,k}$. Therefore, the payment of buyer i, p_i, needs to cover the cost of the model. In other words, the following bid constraint needs to be satisfied:

$$\sum_{j=1}^{N} \left(\sum_{k=1}^{n_j} s_{jk} \left[1 - \prod_{m=1}^{M} \left(1 - \mathcal{I}\left(\beta^i_{j,k,m}\right)\right)\right] + \sum_{l=1}^{L} s_{j,l+n_j} \left[1 - \prod_{m=1}^{M} \left(1 - \mathcal{I}\left(\eta^i_{l,j,m}\right)\right)\right]\right) \leq p_i, \tag{3.8}$$

where $\beta^i_{k,j,m}$ is the coefficient associated to the m^{th} spline variable associated to the k^{th} variable of agent j when predicting $Y_{i,t}$, $\eta^i_{\ell,j,m}$ is the coefficient associated with the m^{th} variable associated to the to the ℓ most recent power measurement from agent j when predicting $Y_{i,t}$, and

$$\mathcal{I}(x) = \begin{cases} 1 & \text{if } x \neq 0 \\ 0 & \text{if } x = 0. \end{cases}$$

If b_i is small, we may not have enough revenue to fulfill the seller's bids $s_{j,k}$.

Therefore, the $\mathcal{SBCLR}$ assumes the form:

$$f(\tilde{\mathcal{Z}}_t; \boldsymbol{\beta}^i, \boldsymbol{\eta}^i) = \beta_0 + \sum_{m=1}^{M} \Bigg\{ \sum_{k=1}^{n_i} \beta^i_{k,i,m} X^k_{i,m,t} + \sum_{j=1, j\neq i}^{N} \sum_{k=1}^{n_j} \beta^i_{k,j,m} X^k_{j,m,t} + \sum_{\ell=1}^{L} \eta^i_{\ell,i,m} Y_{i,m,t} + \sum_{j\neq i} \sum_{\ell=1}^{L} \eta^i_{\ell,j,m} Y_{i,m,t-\ell} + \lambda \left(\sum_{j=1}^{N} \left[\sum_{k=1}^{n_j} \left|\beta^i_{k,j,m}\right| + \sum_{\ell=1}^{L} \left|\eta^i_{\ell,j,m}\right| \right] \right) \Bigg\} \text{subject to}$$
$$\sum_{j=1}^{N} \left(\sum_{k=1}^{n_j} s_{jk} \left[1 - \prod_{m=1}^{M} \left(1 - \mathcal{I}\left(\beta^i_{j,k,m}\right)\right)\right] + \sum_{l=1}^{L} s_{j,l+n_j} \left[1 - \prod_{m=1}^{M} \left(1 - \mathcal{I}\left(\eta^i_{l,j,m}\right)\right)\right]\right) \leq p_i \tag{3.9}$$

Both $\boldsymbol{\beta}^i$ and $\boldsymbol{\eta}^i$ are unknown and must be estimated while aligning with the bid constraint in order to obtain the optimal feasible model that minimizes the lost function among all feasible models. If we choose the loss function $\mathcal{L}\left(\boldsymbol{y}_i, f\left(\tilde{\mathbf{z}}_t; \boldsymbol{\beta}^i, \boldsymbol{\eta}^i\right)\right)$ to be the residual sum of squares (2.6):

$$\hat{\boldsymbol{\beta}}_i, \hat{\boldsymbol{\eta}}_i = \underset{\boldsymbol{\beta},\boldsymbol{\eta}}{\text{argmin}} \frac{1}{2} \sum_{t=1}^{T} \left(\boldsymbol{y}_{i,t} - f\left(\tilde{\mathbf{z}}_t; \boldsymbol{\beta}_i, \boldsymbol{\eta}_i\right)\right)^2 + \lambda \sum_{m=1}^{M} \left(\sum_{j=1}^{N} \left[\sum_{k=1}^{n_j} |\beta^i_{k,j,m}| + \sum_{l=1}^{L} |\eta_{l,j,m}| \right]\right) \text{subject to}$$
$$\sum_{j=1}^{N} \left(\sum_{k=1}^{n_j} s_{jk} \left[1 - \prod_{m=1}^{M} \left(1 - \mathcal{I}\left(\beta^i_{j,k,m}\right)\right)\right] + \sum_{l=1}^{L} s_{j,l+n_j} \left[1 - \prod_{m=1}^{M} \left(1 - \mathcal{I}\left(\eta^i_{l,j,m}\right)\right)\right]\right) \leq p_i \tag{3.10}$$

Due to the budget constraint, there might exist multiple global minimizers of problem (3.10). In such instances, $\boldsymbol{\beta}^i$ and $\boldsymbol{\eta}^i$ are one of those global minimizers. As demonstrated in [26], problem (3.10) can be rewritten as the following problem

$$\begin{aligned} &\min_{\boldsymbol{\Theta}_i} \frac{1}{2} \|\boldsymbol{\Theta}_i - \boldsymbol{a}\|_2^2 + \lambda \sum_{m=1}^{M} \left(\sum_{j=1}^{N} \left[\sum_{k=1}^{n_j} |\beta^i_{k,j,m}| + \sum_{l=1}^{L} |\eta_{l,j,m}| \right]\right) \\ &\text{subject to} \quad \sum_{j=1}^{N} \left(\sum_{k=1}^{n_j} s_{jk} \left[1 - \prod_{m=1}^{M} \left(1 - \mathcal{I}\left(\beta^i_{j,k,m}\right)\right)\right] + \right. \\ &\qquad \left. \sum_{l=1}^{L} s_{j,l+n_j} \left[1 - \prod_{m=1}^{M} \left(1 - \mathcal{I}\left(\eta^i_{l,j,m}\right)\right)\right]\right) \leq p_i, \end{aligned} \tag{3.11}$$

where $\boldsymbol{a} = \frac{1}{TC} \tilde{\mathbf{Z}}^\top \left(\boldsymbol{y}_i - \tilde{\mathbf{Z}}\boldsymbol{\Theta}_i\right)$. This problem is equivalent to solve a knapsack problem.

Proposition 3.1. *If* $\hat{\boldsymbol{\Theta}}_i = (\hat{\boldsymbol{\beta}}^i, \hat{\boldsymbol{\eta}}^i)$ *is an optimal solution to* (3.11)*, then*

$$\hat{\boldsymbol{\Theta}} = sign(\boldsymbol{a} - \lambda) \circ \left(|\boldsymbol{a}| - \lambda\right)_+ \circ \hat{\mathbf{Z}},$$

where $\hat{\mathbf{Z}} = \left(\hat{z}_{1,1}\mathbf{1}_M, \hat{z}_{1,2}\mathbf{1}_M, \ldots, \hat{z}_{jk}\mathbf{1}_M\right)^T$, $\mathbf{1}_M$ *is the row vector of* M *1's, and* $\hat{z}_{1,1}, \hat{z}_{1,2}, \ldots, \hat{z}_{j,k}$ *is the solution to the following 0-1 knapsack problem:*

$$\max_{z_{1,1},z_{1,2},\ldots,z_{jk}} \sum_{j=1}^{N}\sum_{k=1}^{n_j} \underbrace{\left(\sum_{m=1}^{M} \frac{a_{j,k,m}^2 - 2\lambda|a_{j,k,m}| + \lambda^2}{2} \cdot \frac{1 + sign(|a_{j,k,m}| - \lambda)}{2}\right)}_{\mu_{j,k}} z_{j,k} \tag{3.12}$$

$$\text{subject to } \left(z_{1,1}, \ldots, z_{j,k}, \ldots, z_{N,L+n_L}\right) \cdot \left(s_{1,1}, \ldots, s_{j,k}, \ldots, s_{N,L+n_N}\right) \leq p_i.$$

A comprehensive explanation of this procedure can be found in A, while Algorithm 2 enables us to solve a knapsack problem with weights stored in $\boldsymbol{s} \in \mathbb{R}^{p\times 1}$, capacity p_i and item values $\boldsymbol{a} \in \mathbb{R}^{p\times 1}$, where $p = \sum_{j=1}^{N} n_j + NL$.

Algorithm 2 Dynamic Programming for 0-1 Knapsack Problem $\mathcal{KS}$

```
Input: $p$, s, a, $p_i$
Output: Subset of items $I$
for $w = 0$ to $p_i$ do
  $D[0][w] \leftarrow 0$
end for
for $j = 1, \ldots p$ do
  for $w = 0$ to $p_i$ do
    if $s[j] > w$ then
      $D[j][w] \leftarrow D[j-1][w]$
    else
      $D[j][w] \leftarrow \max(D[j-1][w], D[j-1][w - s[j]] + a[j]^2)$
    end if
  end for
end for
$K \leftarrow D[p][p_i]$, $w \leftarrow p_i$, $I \leftarrow \{\}$
for $j \leftarrow p, \ldots, 1$ do
  if $K \leq 0$ then
    break
  end if
  if $K = D[j-1][w]$ then
    continue
  else
    $I \leftarrow I \cup \{j\}$
    $K \leftarrow K - a[j]$
    $w \leftarrow w - s[j]$
  end if
end for
Return $D[p][p_i]$
```

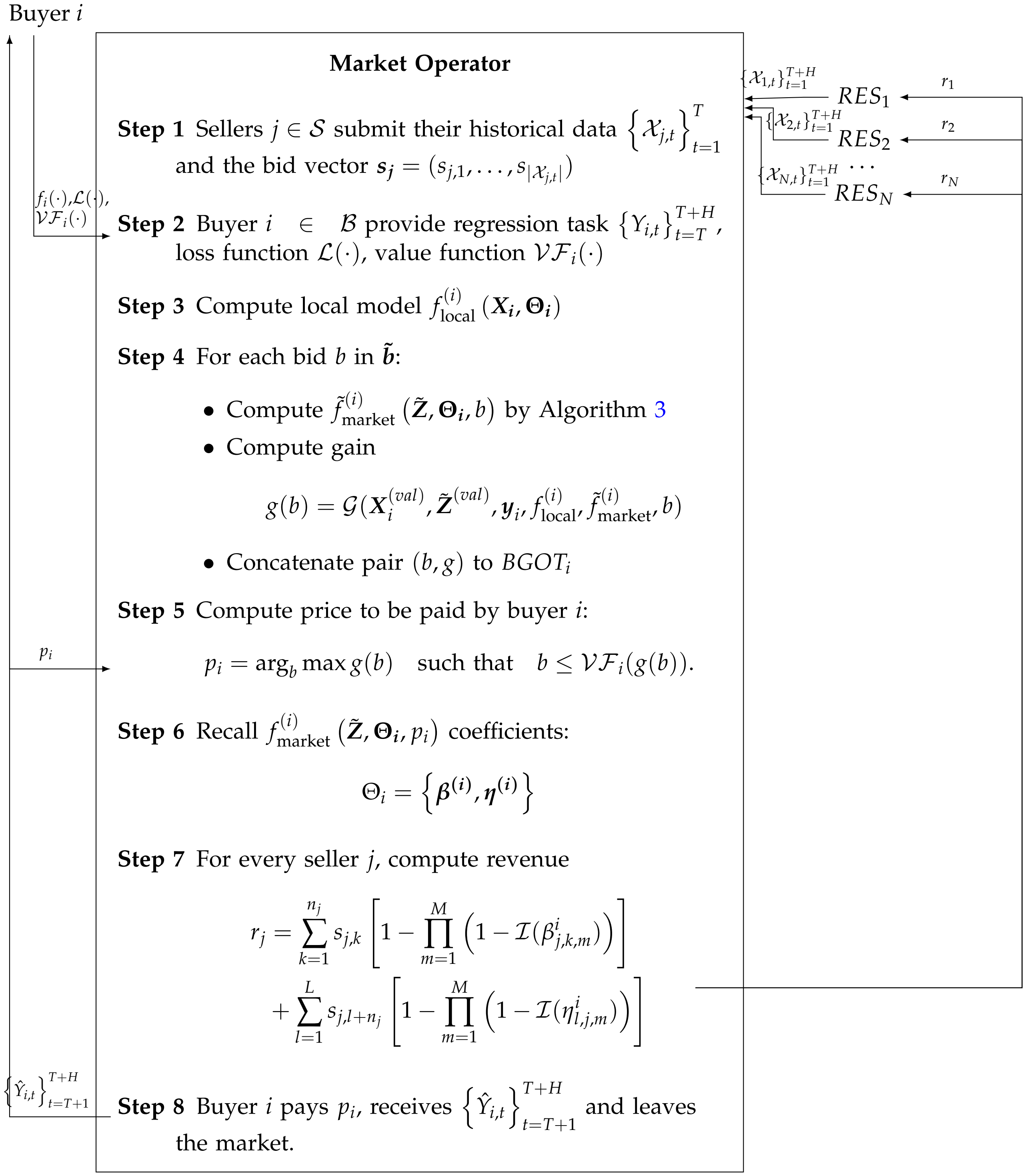
We propose to apply Algorithm 3 to estimate the unknown parameters in $\boldsymbol{\Theta}_i$. Algorithm 3 receives as input the historical data $[\boldsymbol{y}_i, \boldsymbol{Z}]$, the sellers bids $\boldsymbol{s}$, the buyer final price p_i, a positive constant $C > \lambda_{max}(\frac{1}{T}\tilde{\boldsymbol{Z}}^T\tilde{\boldsymbol{Z}})$ (see A for more details), a control parameter δ

and the spline hyperparameters limits $\boldsymbol{\Omega_i} = [\boldsymbol{D}, \boldsymbol{K}]$ where $\boldsymbol{D} = [\boldsymbol{D_{min}}, \boldsymbol{D_{max}}] \in \mathbb{N}^{p \times 2}$, $\boldsymbol{K} = [\boldsymbol{K_{min}}, \boldsymbol{K_{max}}] \in \mathbb{N}^{p \times 2}$ contain the optimal spline order and the optimal number of knots for each feature in $\boldsymbol{Z}$, respectively. We approximate the optimal hyperparameter configuration $\boldsymbol{\Theta_i}$ for our $\mathcal{SBCLR}$ model, to the one of a spline lasso regressor without the bid constraint, computed using Bayesian optimization (Algorithm 1). The function $\mathcal{ST}(\cdot)$ is a spline transformer that operates according to the mechanism explained in Section 2.2.2.

Algorithm 3 Spline Bid-Constrained Lasso Regression $\mathcal{SBCLR}$.

```
**Input:** $\boldsymbol{y}_i, \boldsymbol{Z}, \boldsymbol{s}, p_i, \delta, C, \boldsymbol{\Omega_i}$
**Output:** $\boldsymbol{\Theta}_i \leftarrow (\boldsymbol{\beta}^i, \boldsymbol{\eta}^i)$
$\tilde{\boldsymbol{Z}} \leftarrow \mathcal{ST}(\boldsymbol{Z}, \boldsymbol{\Omega_i})$    ▷ $D, K$ and $\lambda$ estimated through Bayesian Optimization
$\boldsymbol{\Theta}_i^{(0)} \leftarrow (\boldsymbol{\beta}^i, \boldsymbol{\eta}^i)$ such that (3.8) holds
$\Psi \leftarrow \left(\boldsymbol{X}_i, \tilde{\boldsymbol{Z}}, \boldsymbol{y}_i, f_{\text{local}}^{(i)}\right)$
**while** $\mathcal{G}(\Psi, f_{\text{market}}^{(i)}(\boldsymbol{\Theta}_i^{(m)})) - \mathcal{G}(\Psi, f_{\text{market}}^{(i)}(\boldsymbol{\Theta}_i^{(m+1)})) \leq \delta$ **do**
    **for** $iter \geq 1$ **do**
        $\boldsymbol{a}^{(iter)} \leftarrow \boldsymbol{\Theta}_i^{(iter-1)} + \frac{1}{TC}\tilde{\boldsymbol{Z}}^T \left(\boldsymbol{y}_i - \tilde{\boldsymbol{Z}}\boldsymbol{\Theta}_i^{(iter-1)}\right)$
        $\boldsymbol{\mu}_{j,k}^{(iter)} \leftarrow \left(\sum_{m=1}^{M} \frac{(a_{j,k,m}^{(iter)})^2 - 2\lambda|a_{j,k,m}^{(iter)}| + \lambda^2}{2} \cdot \frac{1+\text{sign}(|a_{j,k,m}^{(iter)}| - \lambda)}{2}\right)$
        $\mathbf{w}^{(iter)} \leftarrow \mathcal{KS}(p, \mathbf{s}, \boldsymbol{\mu}^{(iter)}, p_i)$
        $\boldsymbol{\Theta}_i^{(iter)} \leftarrow \boldsymbol{a}^{(iter)} \circ \mathbf{w}^{(iter)} = (\mu_1^{(iter)} w_1^{(iter)}, \mu_2^{(iter)} w_2^{(iter)}, \ldots, \mu_p^{(iter)} w_p^{(iter)})$
    **end for**
**end while**
```

Given a regression task $\{Y_{i,t}\}_{t=T}^{T+H}$, the market operator, to estimate the unknown parameters in $\boldsymbol{\Theta}_i$, performs $\mathcal{SBCLR}$ by solving of the corresponding 0-1 knapsack problem. The research conducted in [26] revealed that the LASSO estimates of $\boldsymbol{\beta_i}$ and $\boldsymbol{\eta_i}$, aligning with the budget constraint while minimizing the average cross-validated error, serve as favorable candidates for $\boldsymbol{\Theta_i}^{(1)}$. However, due to potential local minima in the nonconvex optimization problem (3.10), initiating the algorithm with diverse $\boldsymbol{\Theta_i}^{(1)}$ selections and opting for the solution yielding the least loss function value seems a reasonable strategy. Bayesian optimization (Section 2.2.3) can be used to determine the parameter of regularization λ. By employing this type of regularization, $\mathcal{SBCLR}$ will automatically eliminate the variables that are useless for the regression task at hand.



Market Operator

Step 1 Sellers $j \in \mathcal{S}$ submit their historical data $\left\{\mathcal{X}_{j,t}\right\}_{t=1}^{T}$ and the bid vector $\boldsymbol{s_j} = (s_{j,1}, \dots, s_{|\mathcal{X}_{j,t}|})$

Step 2 Buyer $i \in \mathcal{B}$ provide regression task $\{Y_{i,t}\}_{t=T}^{T+H}$, loss function $\mathcal{L}(\cdot)$, value function $\mathcal{VF}_i(\cdot)$

Step 3 Compute local model $f_{\text{local}}^{(i)}(\boldsymbol{X_i}, \boldsymbol{\Theta_i})$

Step 4 For each bid b in $\tilde{\boldsymbol{b}}$:

- Compute $\tilde{f}_{\text{market}}^{(i)}(\tilde{\mathbf{Z}}, \boldsymbol{\Theta_i}, b)$ by Algorithm 3
- Compute gain

$$g(b) = \mathcal{G}(\boldsymbol{X}_i^{(val)}, \tilde{\boldsymbol{Z}}^{(val)}, \boldsymbol{y}_i, f_{\text{local}}^{(i)}, \tilde{f}_{\text{market}}^{(i)}, b)$$

- Concatenate pair (b, g) to $BGOT_i$

Step 5 Compute price to be paid by buyer i:

$$p_i = \arg_b \max g(b) \quad \text{such that} \quad b \leq \mathcal{VF}_i(g(b)).$$

Step 6 Recall $f_{\text{market}}^{(i)}(\tilde{\boldsymbol{Z}}, \boldsymbol{\Theta_i}, p_i)$ coefficients:

$$\Theta_i = \left\{\boldsymbol{\beta}^{(i)}, \boldsymbol{\eta}^{(i)}\right\}$$

Step 7 For every seller j, compute revenue

$$r_j = \sum_{k=1}^{n_j} s_{j,k}\left[1 - \prod_{m=1}^{M}\left(1 - \mathcal{I}(\beta_{j,k,m}^{i})\right)\right] + \sum_{l=1}^{L} s_{j,l+n_j}\left[1 - \prod_{m=1}^{M}\left(1 - \mathcal{I}(\eta_{l,j,m}^{i})\right)\right]$$

Step 8 Buyer i pays p_i, receives $\left\{\hat{Y}_{i,t}\right\}_{t=T+1}^{T+H}$ and leaves the market.

Buyer i
$f_i(\cdot)$,$\mathcal{L}(\cdot)$,
$\mathcal{VF}_i(\cdot)$
p_i
$\left\{\hat{Y}_{i,t}\right\}_{t=T+1}^{T+H}$
$\{\mathcal{X}_{1,t}\}_{t=1}^{T+H}$
RES_1
r_1
$\{\mathcal{X}_{2,t}\}_{t=1}^{T+H}$
RES_2
r_2
$\cdots$
$\{\mathcal{X}_{N,t}\}_{t=1}^{T+H}$
RES_N
r_N

FIGURE 3.2: Auction Data Market mechanism at time $t = T$.

3.4 Data Market Mechanism

The data market mechanism operates systematically, facilitating the exchange of information between sellers and buyers. This process is defined by a sequence of well-defined steps catering to both parties' needs.

Figure 3.2 summarizes the proposed mechanism. The process begins at a specific time $t = T$ when the market operator opens a session. During this phase, sellers $j \in \mathcal{S}$ submit their historical data $\left\{\mathcal{X}_{j,t}\right\}_{t=1}^{T}$ and bid vectors $\boldsymbol{s_j} = (s_{j,1}, \dots, s_{j,n_j+L})$ to the market operator. Simultaneously, buyers $i \in \mathcal{B}$ provide their regression tasks $\left\{Y_{i,t}\right\}_{t=T}^{T+H}$ and their value functions $\mathcal{VF}_i$.

Following this initial stage, a closed session ensues. In this phase, the market operator handles all regression tasks in parallel. For each submitted task, the following steps unfold. Firstly, the market operator performs the local model $f^i_{local}(\boldsymbol{X_i}, \boldsymbol{\Theta_i})$ using the forecasting model chosen by the buyer i, trained with historical data $\mathbf{X}_i$. Subsequently, for each bid b in $\tilde{\boldsymbol{b}}$, the market $\mathcal{SBCLR}$ model $f^{(i)}_{\text{market}}(\mathbf{Z}, \boldsymbol{\Theta_i})$ is computed, and its gain function values are evaluated using the loss function $\mathcal{L}(\cdot)$ given by buyer i, in order to establish the relation between forecasting skill enhancement and bids. This information is presented in $BGOT_i$. With this table, the market operator computes the final price to be payed by agents i given by:

$$p_i = \arg_b \max g(b) \quad \text{such that} \quad b \leq \mathcal{VF}_i(g(b)). \tag{3.13}$$

The market recalls the market coefficients $\Theta_i = \left\{\boldsymbol{\beta}^{(i)}, \boldsymbol{\eta}^{(i)}\right\}$ associated with the price p_i in order to be able to compute the revenues for each seller j that are given by

$$r_j = \sum_{k=1}^{n_j} s_{j,k} \left[1 - \prod_{m=1}^{M} \left(1 - \mathcal{I}(\beta^i_{j,k,m})\right)\right] + \sum_{l=1}^{L} s_{j,l+n_j} \left[1 - \prod_{m=1}^{M} \left(1 - \mathcal{I}(\eta^i_{l,j,m})\right)\right]$$

Finally, the buyer i pays p_i, receives $\left\{Y_{t,i}\right\}_{t=T+1}^{T+H}$, marking the conclusion of their engagement in the market.

Note that by incorporating a LASSO Regressor into the market framework, we are performing feature selection regulated by the importance of the features for the regression task at hand and by the sellers bids. On the one hand, sellers receive always what they bid for the usage of their data. However if they bid their data to a high value, their data may never be used and they don't generate any revenue from the market. If different agents have similar data, their participation in a given forecasting task will be determined by their bids.

On the other hand, the market allows buyers to express their willingness to pay accordingly to the gain, making them only pay what they bid.

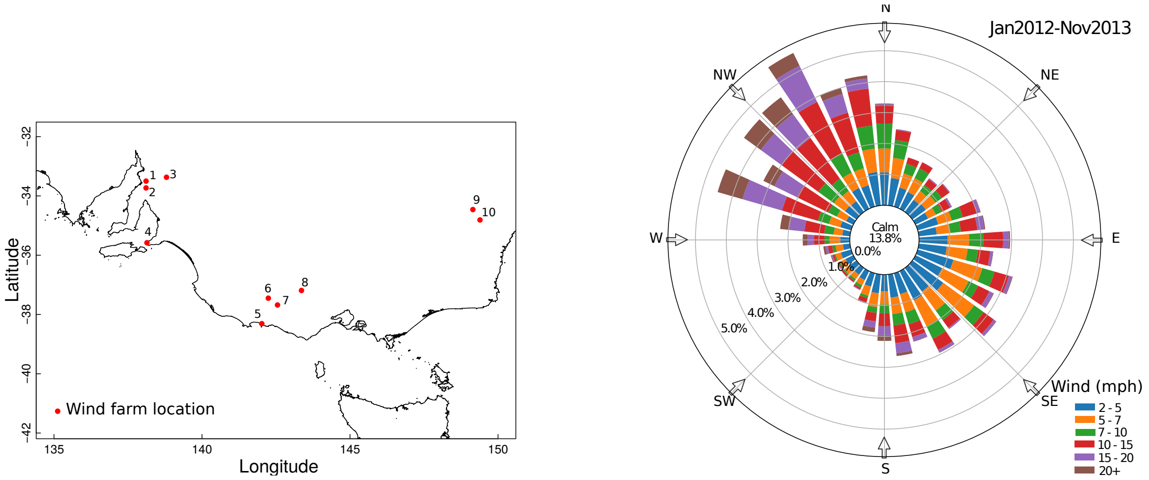
In summary, our framework aligns market dynamics with data value and economic interests, promoting fairness, efficiency, and collaboration in energy sector forecasting, while satisfying both buyers and sellers interests.

Chapter 4

Case Study

4.1 Problem and data description

This case study aims to predict, at midnight, the generation of wind power 24 hours ahead for ten different zones in Australia, as depicted in Figure 4.1a. Each zone corresponds to a specific wind farm. The wind power measurements were normalized based on the nominal capacity of each wind farm. This dataset was originally employed in the Global Energy Forecasting Competition 2014 (GEFCom2014) and is publicly available. It encompasses the period from January 1, 2012, to November 30, 2013.



(A) Wind farm's location (extracted from [6])

(B) Wind rose (Canberra: close to Zone 9) (extracted from [6])

FIGURE 4.1: GEFCom2014 wind power dataset.

The predictors include wind forecasts at two heights, 10 and 100 m above ground level, obtained from the European Centre for Medium-range Weather Forecasts (ECMWF). These forecasts were for the zonal and meridional wind components (denoted U and V), i.e., projections of the wind vector on the west-east and south-north axes, respectively.

The ECMWF provided weather forecasts to the precise locations of the wind farms, issued daily at midnight, offering an hourly resolution for a 24-hour projection ahead. These forecasts are aligned with the specifications of the forecasting exercise and were used as inputs for training and evaluating the various forecast models.

It is worth noting that these weather forecasts are not only available for the training phase but also serve as essential inputs for the multifaceted tasks employed in the evaluation of our forecasting methodologies. Moreover, in the context of the training data provided, we are furnished with power measurements obtained from various wind farms. These measurements are recorded hourly but are limited to the training period.

We map this forecasting task to the RES Forecasting problem described in Section 2.1. The data was collected from 10 different zones, meaning there are $N = 10$ agents. Here we consider that each agent acts as a *data prosumer*, i.e., a data owner that consume and supply data to the market ($\mathcal{A} = \mathcal{B} = \mathcal{S}$). Each agent $i \in \mathcal{A} \in \{1, 2, \dots, 10\}$ possesses four exogenous variables $X^1_{i,t}, X^2_{i,t}, X^3_{i,t}$ and $X^4_{i,t}$ which correspond to the zonal and meridional wind components at 10 and 100m above, and one endogenous variable $Y_{i,t}$ which represent the power measurements.

A rolling basis approach is adopted to make forecasts: a sliding window of one month test is used, and the model's fitting period encompasses 12 months. Consequently, for each zone, each model is optimized 11 times.

4.2 Data analysis and feature engineering

In this section, we aim to comprehensively explore the ECMWF dataset to understand the dynamics between weather measurements and power generation among RES agents.

4.2.1 Data analysis

The wind power output series from the wind farms are shown in Figure 4.2. These time series plots show how power generation varies over time, between 2012 and 2013, where one can not see significant patterns of seasonalities or trends.

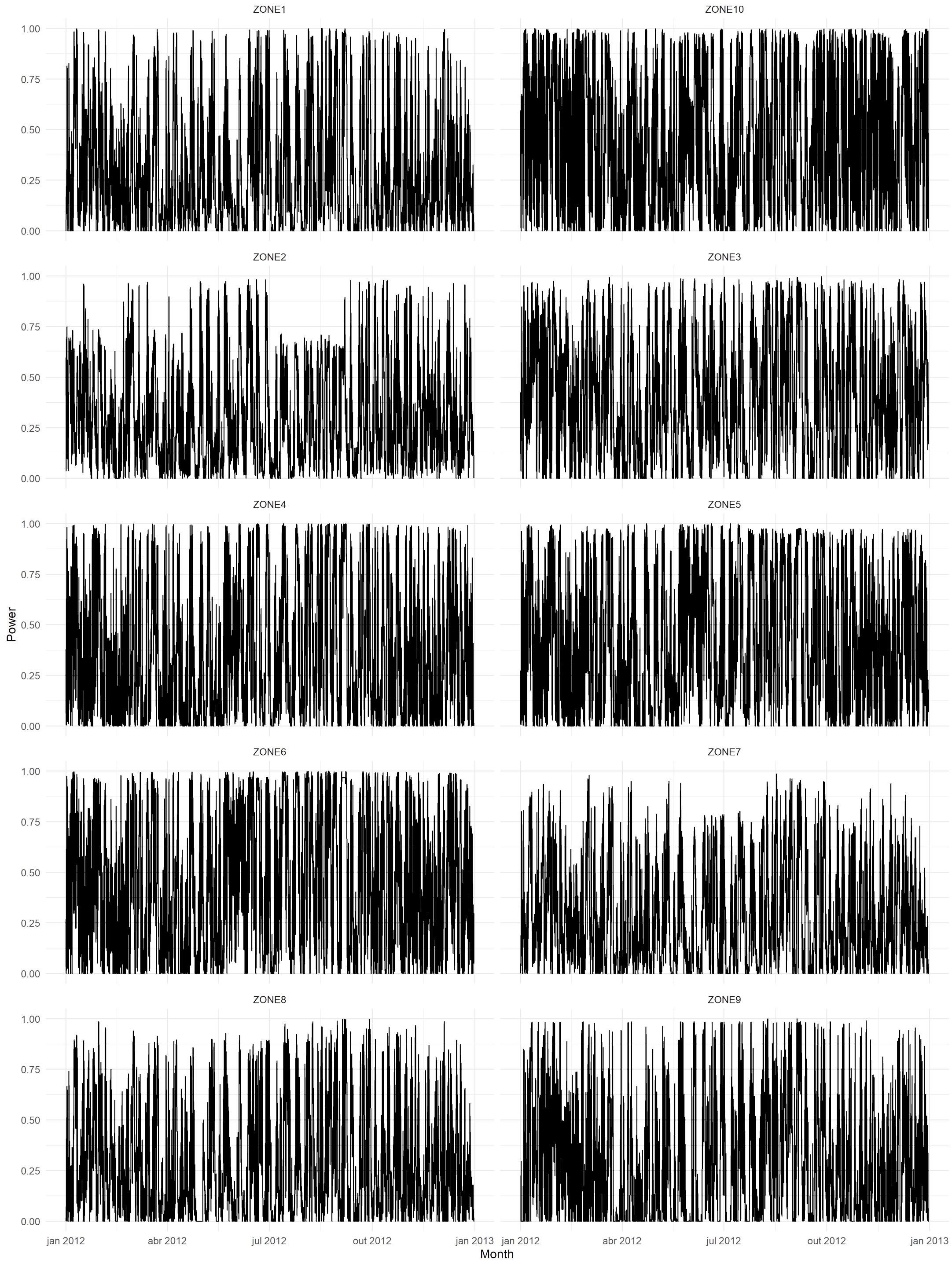


FIGURE 4.2: Wind power generation in GEFCom2014

The exogenous variables relations with the power generation at wind farm 1 are shown in Figure 4.3 where one can not see any linear dependencies, calling the need for forecasting methodologies that can capture these patterns.

In the literature, several studies have shown that RES agents can exhibit intricate spatio-temporal dependencies. These dependencies refer to how different RES agents,

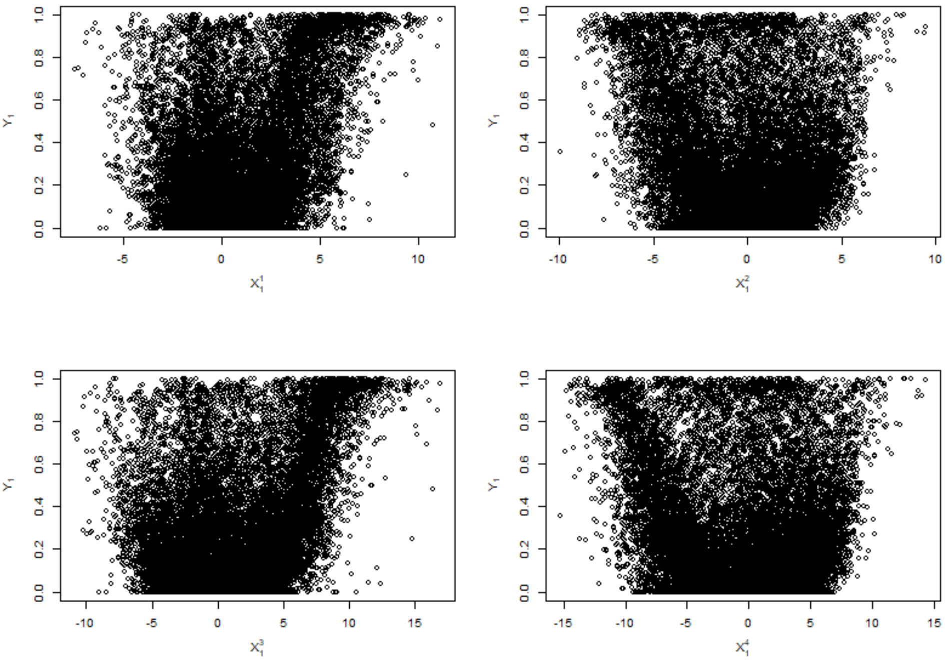


FIGURE 4.3: Scatter matrices of wind power generation and weather forecasts for wind farm 1 in GEFCom2014.

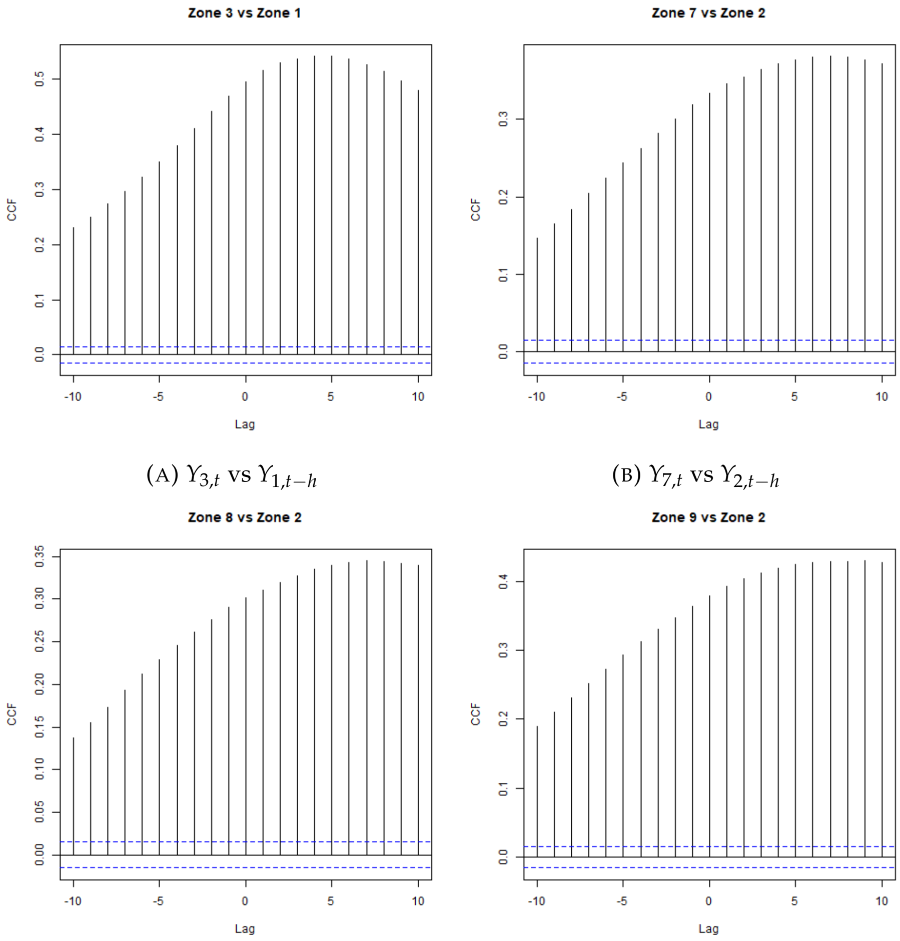
such as wind power plants, solar panels, or other renewable energy sources, are interconnected not only across space but also over time. In simpler terms, it is about understanding how the behavior of one RES agent at a particular location and moment in time can influence or be influenced by other agents located elsewhere or at different times. To address if these spatio-temporal dependencies observed in the literature indeed manifest in the GEFCom2014 dataset we employ a cross-correlation analysis.

Cross-correlation is a statistical technique to measure the similarity or relationship between two different time series or signals. It helps us understand how one time series is related to another when shifted or lagged in time. Mathematically, the cross-correlation between the power measurements of wind farm $i \in \mathcal{A}$ with the h-lagged power measurements of wind farm $j \in \mathcal{A}$ is defined as:

$$CCF_{i,j}(h) = \frac{\sum_{t=1}^{T} \left(Y_{i,t} - \bar{Y}_i\right) \left(Y_{j,t+h} - \bar{Y}_j\right)}{\sqrt{\left(\sum_{t=1}^{T} \left(Y_{i,t} - \bar{Y}_i\right)^2\right) \left(\sum_{t=1}^{T} \left(Y_{j,t} - \bar{Y}_j\right)^2\right)}} \tag{4.1}$$

where $\bar{Y}_i = \frac{1}{T} \sum_{t=1}^{T} Y_{i,t}, \quad i \in \mathcal{A}.$

In our study, we have observed that spatio-temporal information becomes increasingly relevant for longer lead-times, as evidenced by the cross-correlation plots presented in Figure 4.4.



(A) $Y_{3,t}$ vs $Y_{1,t-h}$

(B) $Y_{7,t}$ vs $Y_{2,t-h}$

(C) $Y_{8,t}$ vs $Y_{2,t-h}$

(D) $Y_{9,t}$ vs $Y_{2,t-h}$

FIGURE 4.4: Cross-correlation plots

These plots illustrate the cross-correlations among a selection of wind power plants. Notably, the cross-correlation values continue to rise until reaching a lag of 6 hours. This observation implies that, for instance, the current power generation at Zone 9 exhibits a stronger correlation with the power generation at Zone 2 from 6 hours prior. This phenomenon aligns with our expectations, as it can be attributed to the geographical distribution of the wind farms, as illustrated in Figure 4.1a. The region's meteorological characteristics, particularly wind speed, contribute to these correlations. Figure 4.1b illustrates the wind rose diagram for a location near Zone 9. It reveals that, over the two years under consideration, wind direction exhibited considerable variability. However, the most robust and consistent winds predominantly originated from the northwest or west directions. This pattern indicates that wind power plants situated to the east (e.g., Zone 9, Zone 10) or southeast (e.g., Zone 5, Zone 6, Zone 7, Zone 8) can benefit significantly from leveraging the lagged data from wind farms located to the west (Zone 1 to Zone 4).

4.2.2 Feature engineering

In the context of our RES forecasting problem, we aim to predict wind power generation for each hour of the day D, starting from 01:00 to 24:00, where day D at 24:00 effectively corresponds to the following day, day $D+1$ at 00:00.

One of the critical questions that naturally arises in this context is determining the appropriate number of lags, denoted as L, to be incorporated into our forecasting models. Based on correlation analysis and previous experiences with the dataset, it was observed that a significant relationship between power generation and its historical values exists up to a lag of 6 hours. Consequently, our strategy integrates the most recent six lagged power measurements as essential features for predictions within the 1 to 6-hour forecasting horizon.

However, it is essential to consider the practicality of this approach in a real-world scenario. When we are forecasting power generation only 1 hour into the future ($H = 1$), our feature set includes the lagged values of all power measurements, denoted as $Y_{j,t-1}$ through $Y_{j,t-6}$, $\forall j \in \mathcal{A}$, and the exogenous variables of agent i, $X^1_{i,t}, X^2_{i,t}, X^3_{i,t}$ and $X^4_{i,t}$.

However, we encounter a practical constraint as we extend our forecasting horizon to 2 hours into the future ($H = 2$). At the forecasting time point of 2:00, data for the 1-hour lag ($Y_{i,t-1}$) is not available due to the timing of our forecasts, all of which originate at midnight (00:00). Consequently, for 2-hour ahead predictions, our feature set is adjusted to encompass only the lagged time series $Y_{i,t-2}$ through $Y_{i,t-5}$, $\forall j \in \mathcal{A}$, ensuring that our models are built on available historical data. The exogenous variables $X^1_{i,t}, X^2_{i,t}, X^3_{i,t}$ and $X^4_{i,t}$ are also included. For forecasting 3 hours ahead, our feature set includes the lagged power measurements $Y_{i,t-3}$ through $Y_{i,t-5}$, $\forall j \in \mathcal{A}$.

For predictions beyond the 6-hour horizon (hours 7 and beyond), the relationship between power and historical observations diminishes, and the focus shifts primarily to weather forecasts. Therefore, for hours 7, 8, 9, and so forth, a unified model can be employed, relating forecasts of wind components $X^1_{i,t}, X^2_{i,t}, X^3_{i,t}$ and $X^4_{i,t}$ with power generation $Y_{i,t}$.

Table 4.1 illustrates which lagged power measurements are incorporated into our models according to each forecasting horizon, when the forecasting task is instantiated at 00:00 of day D. This meticulous feature selection aligns with the accessibility of historical data at the specific time of prediction, ensuring that our forecasting models are optimized for accuracy and effectiveness across various forecasting horizons.

Target	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6
D+1 01:00	D+1 00:00	D 23:00	D 22:00	D 21:00	D 20:00	D 19:00
D+1 02:00	–	D+1 00:00	D 23:00	D 22:00	D 21:00	D 20:00
D+1 03:00	–	–	D+1 00:00	D 23:00	D 22:00	D 21:00
D+1 04:00	–	–	–	D+1 00:00	D 23:00	D 22:00
D+1 05:00	–	–	–	–	D+1 00:00	D 23:00
D+1 06:00	–	–	–	–	–	D+1 00:00
D+1 H:00, $H > 7$	–	–	–	–	–	–

TABLE 4.1: Lags used for each forecasting horizon

4.3 Comparison of forecasting models

In this section, we delve into the comprehensive comparison of various forecasting models to gain insights into their performance and applicability in the context of wind power generation prediction. The models under scrutiny include Linear Regression, Lasso Linear Regression, Spline Linear Regression, Spline Lasso Regression, and the Gradient Boosting Regressor, all explained in Section 2.2.2. The selection and fine-tuning of hyperparameters were done using Bayesian Optimization, as explained in Algorithm 1, and the loss function adopted was the $RMSE$. Table 4.2 provides an overview of the evaluated models and their corresponding hyperparameter search ranges.

A rolling basis approach is adopted to make forecasts: a sliding window of one month test is used, and the model's fitting period encompasses 12 months. Consequently, each model is optimized 11 times for each zone, including the hyperparameters. We evaluate the performance of these models by assessing the Root Mean Squared Error ($RMSE$) values across different hours of the day for each of the ten distinct wind farm zones. These $RMSE$ values were derived by averaging the results over all 11 months within each zone. Each subplot in Figures 4.5a and 4.5b represent the $RMSE$ and MAE (per hour) for the collaborative models with the data construction explained in Section 4.2.2 depicted with continuous lines, and models using only local data (agent's own models), represented by dashed lines.

TABLE 4.2: Evaluated models and tuned parameters.

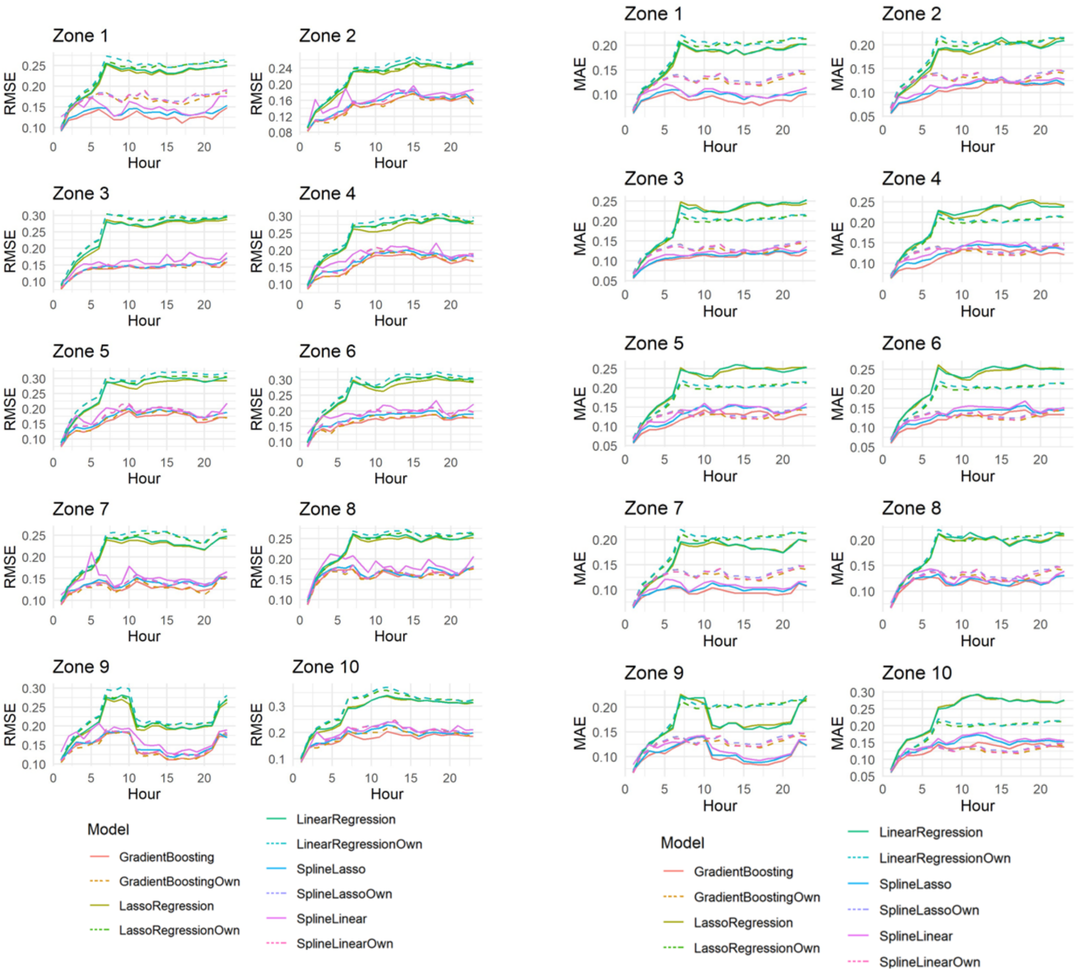
Model	Hyper-parameters	Range
Linear	–	–
Lasso Linear	Regularization (λ)	$[10^{-3}, 1]$
Linear Spline	B-splines degree (D)	$\{1, \ldots, 7\}$
	Number of knots (K)	$\{10, \ldots, 30\}$
Lasso Spline	Regularization (λ)	$[10^{-3}, 1]$
	B-splines degree (D)	$\{1, \ldots, 7\}$
	Number of knots (K)	$\{10, \ldots, 30\}$
Gradient Boosting	Maximum depth (max_depth)	$\{3, \ldots, 10\}$
	Learning rate (η)	$[10^{-3}, 1]$
	Maximum number of features ($max\_features$)	$\{1, \ldots, 900\}$
	Minimum number of samples required to split an internal node ($min\_samples\_split$)	$\{10, \ldots, 50\}$
	Minimum number of samples by leaf node ($min\_samples\_leaf$)	$\{10, \ldots, 50\}$
	Fraction of samples to be used for fitting the individual base learners ($subsample$)	$[0.7, 0.9]$
	Number of boosting stages to perform ($n\_estimators$)	$\{50, \ldots, 100\}$

One prominent trend was the consistent superiority of collaborative models over local models in terms of RMSE. Collaborative models consistently outperformed their local counterparts, although the extent of this advantage varied across zones. This underscored the potential benefits of collaboration and information sharing among zones, with the degree of improvement varying geographically.

The Gradient Boosting Regressor emerged as a frontrunner among the models evaluated, consistently delivering lower RMSE values. Remarkably, the Spline Lasso Regression model, a linear parameter model, closely matched the performance of Gradient Boosting. This suggests that additive linear models can be effective alternatives in energy forecasting.

In the MAE analysis, the stability of collaborative models' differences across different hours of the day contrasted with the more heterogeneous distribution observed in $RMSE$ plots. Smaller differences in MAE were also noted, aligning with our optimization focus on RMSE, which emphasizes larger errors.

The $RMSE$ and MAE analysis were the average values across the 11 months. This prevent us to see if the collaboration was beneficial in all 11 months or in only in some. To evaluate if there is actually a significant difference between the collaborative and local models using a spline lasso regression and gradient boosting regression in all months,



(A) *RMSE* of collaborative and local machine learning models averaged out along the 11 testing months.

(B) *MAE* of collaborative and local machine learning models.

FIGURE 4.5: Comparison of RMSE and MAE

we employed the Diebold-Mariano hypothesis test, a statistical tool designed to compare the forecast accuracy of two distinct methods [27]. Let $\hat{Y}^{\text{market}}_{i,t}=f^{(i)}_{\text{local}}(\mathcal{X}_{i,t};\Theta_i)$ and $\hat{Y}^{\text{local}}_{i,t}=f^{(i)}_{\text{market}}(\mathcal{Z}_t;\Theta_i)$ be the forecasting series for the market collaborative model and for the agent's own local model, respectively. Supposing the forecasting errors are $e^{\text{market}}_{i,t} = Y_{i,t} - \hat{Y}^{\text{market}}_{i,t}$ and $e^{\text{local}}_{i,t} = Y_{i,t} - \hat{Y}^{\text{local}}_{i,t}$, the accuracy of each forecast is measured by a function $\mathcal{L}$ which in our case is the *RMSE* loss function. The equal accuracy hypothesis is tested to determine if there is a significant statistical difference between the market and local models. Mathematically, the null hypothesis is

$$\text{H}_0 : \mathbb{E}\left[d_{i,t}\right] = 0,$$

where $d_{i,t} = \mathcal{L}\left(e_{i,t}^{\text{market}}\right) - \mathcal{L}\left(e_{i,t}^{\text{local}}\right)$, and the alternative hypothesis is usually

$$\mathrm{H}_1 : \mathbb{E}\left[d_{i,t}\right] \neq 0.$$

By employing the $RMSE$, the empirical value for $\mathbb{E}\left[d_{i,t}\right]$ is the sample mean

$$\bar{d}_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(e_{i,t}^{\text{local}}\right)^2} - \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(e_{i,t}^{\text{market}}\right)^2}.$$

The sample mean, $\bar{d}_i$, converges asymptotically to a normal distribution under the condition that the loss differential is a covariance stationary series and the DM test statistic is

$$\mathrm{DM} = \frac{\bar{d}_i}{\sqrt{\frac{2\pi\hat{f}_d(0)}{T}}} \longrightarrow \mathcal{N}(0,1).$$

Here, $2\pi\hat{f}_d(0)$ represents a consistent estimate of the asymptotic variance of $\sqrt{T}\bar{d}$ based on sample autocovariance.

To assess the statistical significance of the DM test results, a common significance level of 5% is typically adopted. Under this threshold, the null hypothesis is rejected if the absolute value of DM exceeds 1.96. This rejection indicates that there is a significant difference between the forecast accuracy of the two models being compared.

In our case, we sought to discern whether collaborative market models exhibited lower $RMSE$ values, and hence higher forecast accuracy, when juxtaposed against local models. The DM test can also be applied to test the null hypothesis against the alternative hypothesis that the collaborative market model performs better than the local model, $\mathrm{H}_1 : \mathbb{E}\left[d_{i,t}\right] > 0$. The focus here is on the alternative hypothesis $\mathrm{H}_1 : \mathbb{E}\left[d_{i,t}\right] > 0$, i.e., local model errors were greater, meaning their inferior forecasting capabilities. In this case, the null hypothesis is rejected for a significance level of 5% if $\mathrm{DM} > 1.64$ [28].

Our examination is extended to varying hours of the day, assessing each set and zone individually. The outcomes unveiled that different levels of leverage of participation in the data market discriminated in Table 4.3 that show us the percentage of cases that the alternate hypothesis $\mathrm{H}_1 : \mathbb{E}\left[d_{i,t}\right] > 0$ verifies with a significance level of 5%, for the models Spline Lasso Regressor and Gradient Boosting Regressor, respectively.

It is paramount to underscore that our overarching analysis of the two models with higher performance revealed that collaborative models yielded enhanced forecasting skill. This was contingent upon the assumption of zero loss in market participation.

TABLE 4.3: Percentage of testing months (out of 11) where H_0 is rejected with a significance level of 5%.

(A) Spline Lasso Regressor

Zone	Percentage (%)
1	100.0
2	27.27
3	27.27
4	36.36
5	63.64
6	63.64
7	45.45
8	36.36
9	27.27
10	81.82

(B) Gradient Boosting Regressor

Zone	Percentage (%)
1	100.0
2	18.18
3	0.0
4	36.36
5	63.64
6	45.45
7	9.09
8	9.09
9	18.18
10	63.64

We present a summary of the average computational times for each model in Table 4.4. This computational time encapsulates the training phase of the model as well as the computation of the forecasts. The Gradient Boosting Regressor consumes the most time for generating predictions. In contrast, the Spline Lasso Regression model executes the forecasting tasks more than four times faster than the Gradient Boosting model. This underscores the practicality and feasibility of the Spline Lasso Regression model, which retains competitive forecasting capabilities while minimizing computational time.

TABLE 4.4: Evaluated models and tuned parameters.

Model	Average time (s)
Linear	0.04
Lasso Linear	16.01
Linear Spline	50.88
Lasso Spline	93.30
Gradient Boosting	414.10

4.4 Data market simulation

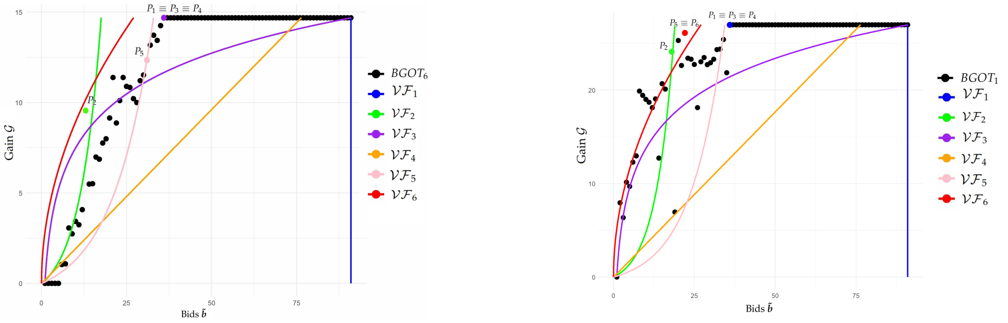
We performed a simulation of our data market proposal using the GEFCom2014 dataset. In our simulation, we assumed that all local models ($\mathcal{M}_i$) utilized a Spline Lasso Regression approach. These models were trained using local data, and hyperparameters K, D, and λ were fine-tuned through Bayesian optimization. We chose a consistent loss function for all buyers, specifically $\mathcal{L}_i(\cdot) = \text{RMSE}(\cdot)$.

While in reality, seller bids may vary, for our experiment, we set all seller bids to a uniform unit price, considering simplicity ($s_{j,k} = 1$). In the context of Algorithm 3, which governs the regression task for each agent i, we initialized the regression coefficients $\mathbf{\Theta}_i^{(0)}$ by equating the coefficients of the local variables $\mathcal{X}_{i,t}$ to those of their respective local model $\mathcal{M}_i$, while setting all other coefficients to zero.

Concerning the pricing determination process, we explored seven distinct value functions ($\mathcal{VF}_1$ to $\mathcal{VF}_6$):

- $\mathcal{VF}_1(g) = \max(\tilde{\boldsymbol{b}})$
- $\mathcal{VF}_2(g) = 0.5 \cdot \frac{\max(\tilde{\boldsymbol{b}})}{\ln(1+3\cdot\max(\boldsymbol{g}))} \cdot \ln(1 + 3 \cdot \max(\boldsymbol{g})) \cdot g$
- $\mathcal{VF}_3(g) = e^{\frac{\ln(\max(\tilde{\boldsymbol{b}}))}{\max(g)} \cdot g}$
- $\mathcal{VF}_4(g) = \frac{\max(\tilde{\boldsymbol{b}})\cdot 0.85 - \min(\tilde{\boldsymbol{b}})}{\max(\boldsymbol{g}) - \min(\boldsymbol{g})} \cdot g$
- $\mathcal{VF}_5(g) = 0.35 \frac{\max(\tilde{\boldsymbol{b}})}{\ln(1+10\cdot\max(\boldsymbol{g}))} \cdot \ln(1 + 10 \cdot \max(\boldsymbol{g})) \cdot g$
- $\mathcal{VF}_6(g) = 0.30 \cdot \frac{(\max(\tilde{\boldsymbol{b}}) - \min(\tilde{\boldsymbol{b}}))}{(\max(\boldsymbol{g}) - \min(\boldsymbol{g}))^2} \cdot g^2$

These functions were meticulously tuned to align with the scale of gains achieved in the forecasting endeavors, $\boldsymbol{g}$, and with potential bids $\tilde{\boldsymbol{b}} = [1, 91]$. It is important to emphasize that buyers would have no insight into potential gains in a real-world scenario, necessitating a strategic approach to tune their value functions for maximum benefit. To estimate the gain, we calculated the $RMSE$ values using a single fold of size $\Delta = 31 \times 24$, corresponding to 1 month. In Figures 4.6a and 4.6b we illustrate two examples of price definition according to (3.5), where points P_1 through P_6 represent the pivotal determinants of the price to be paid and the corresponding estimated gain computed according to (3.3). It is evident that, after a certain point, regardless of the buyer's willingness to pay, the gain from forecasting skill does not change significantly. This signifies a saturation point where all valuable data has been leveraged. Any further variable selections would have minimal impact on forecasting skill gain.



(A) Value functions and $BGOT_6$ for the predictions in January 2013

(B) Value functions and $BGOT_1$ for the predictions in April 2013

FIGURE 4.6: Illustrations of price definition.

In Table 4.5, we present an analysis of cumulative gains observed during the year 2013 for different value functions. The mean cumulative gains, arranged in descending order across various value functions, are as follows: Buyer 1 $>$ Buyer 10 $>$ Buyer 6 $>$ Buyer 5 $>$ Buyer 4 $>$ Buyer 9 $>$ Buyer 2 $>$ Buyer 7 $>$ Buyer 8 $>$ Buyer 3. This ranking aligns consistently with the findings presented in Table 4.3 for the Spline Lasso Regressor.

Examining the percentages of cases where agents benefited from using market data, as shown in Table 4.3, we observe the following descending order: Buyer 1 $>$ Buyer 10 $>$ Buyers 5 and 6 $>$ Buyer 7 $>$ Buyers 4 and 8 $>$ Buyers 2, 3, and 9. It is evident that Buyer 1 derived the most significant forecasting skill improvement from utilizing market data. Notably, in Table 4.3, Zone 1 consistently demonstrated a 100% benefit rate from market data,rejecting hypothesis H_0 of equal performance.

Furthermore, Buyer 10 emerges as the second most proficient agent in extracting gains from the market, coinciding with the higher percentage observed in Table 4.3, where Buyer 10 also ranks second.

Table 4.6 illustrates the cumulative payments made by each buyer for all value functions. It is evident that the buyer who paid the highest amount is Buyer 1, corresponding to the one who obtained the most significant gains from the market. Notably, the order of payments mirrors the order of gains obtained, highlighting a direct relationship between the amount paid and the gains achieved.

Figure 4.7 illustrates the cumulative gains observed during the year 2013 when employing the value function $\mathcal{VF}_1$. A noticeable trend emerges as we observe the behavior of both curves, which appear to closely mirror each other. However, there are instances where the market underestimates gains, as evident in Zone 9, while in other cases, it tends

TABLE 4.5: Cumulative Gains (%)

$\mathcal{B}$	$\mathcal{VF}_1$	$\mathcal{VF}_2$	$\mathcal{VF}_3$	$\mathcal{VF}_4$	$\mathcal{VF}_5$	$\mathcal{VF}_6$	**Mean**
Buyer 1	2926.58	2269.42	1304.83	750.02	2995.42	2995.42	2197.41
Buyer 2	446.15	375.50	289.47	180.47	500.31	500.31	365.13
Buyer 3	171.11	85.86	69.66	47.51	185.21	185.21	117.95
Buyer 4	710.76	554.54	277.57	196.00	804.15	804.15	487.68
Buyer 5	778.97	560.89	186.70	105.51	882.80	882.80	573.43
Buyer 6	1051.56	874.33	298.78	170.03	1117.96	1117.96	764.57
Buyer 7	352.81	256.18	48.49	9.39	368.01	368.01	253.00
Buyer 8	337.81	227.64	65.10	27.42	341.31	341.31	235.16
Buyer 9	515.52	456.71	286.73	118.75	515.52	515.52	378.79
Buyer 10	1125.46	957.62	326.68	463.58	1136.00	1136.00	813.89
Mean	760.68	573.68	286.21	259.58	831.21	831.21	

TABLE 4.6: Cumulative Payments at the End of 1 Year

$\mathcal{B}$	$\mathcal{VF}_1$	$\mathcal{VF}_2$	$\mathcal{VF}_3$	$\mathcal{VF}_4$	$\mathcal{VF}_5$	$\mathcal{VF}_6$	**Mean**
Buyer 1	9753	4948	9753	8531	1128	2208	5816.50
Buyer 2	3362	1579	3362	2742	445	1285	2253.33
Buyer 3	1666	391	1666	1247	90	345	830.83
Buyer 4	5503	2645	5503	4031	461	1086	2760.50
Buyer 5	6407	3132	6407	5191	332	1154	3035.50
Buyer 6	6753	4078	6753	5701	490	1726	3536.50
Buyer 7	3405	1468	3405	2900	59	446	1752.17
Buyer 8	3043	1238	3043	2850	108	535	1639.50
Buyer 9	3476	2576	3476	3476	318	1722	2370.33
Buyer 10	5023	3047	5023	4790	1420	1389	3431.67
Sum	48391	25102	48391	41459	4851	11896	

to overestimate them, as exemplified in Zone 2. In fact, the distribution of the gains estimated and observed in the course of 1 year as very similar for most zones as depicted in Figure 4.8. To analyze an agent's typical performance when generating hourly forecasts throughout a day, we calculated both the average gain and the average price per hour. Assuming that the agent opts for a function value of $\mathcal{VF}_1$, Figures 4.9a and 4.9b illustrates the cumulative trends for both the average gain and the average price, respectively, over the course of 24 hourly forecasts within a single day. We can see that, on average, it is possible to achieve an accumulated gain of 80% for hourly forecasts in one day while paying less than 440 price units. Having demonstrated the substantial gains a buyer can achieve by engaging in the market over the course of a day, we now shift our focus to elucidate the advantages for sellers. Table 4.7 showcases the cumulative revenues generated by sellers throughout the year 2013. Notably, it becomes evident that Zone 2 emerges as the

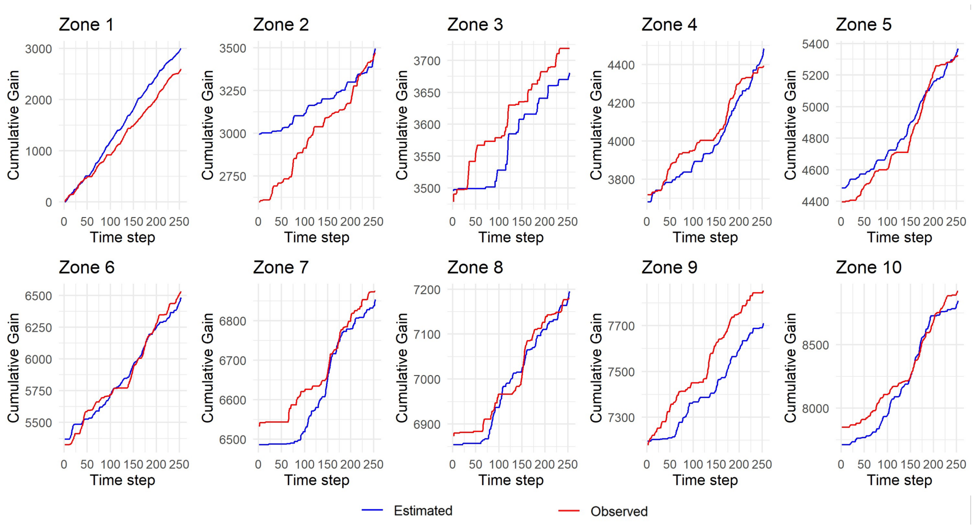


FIGURE 4.7: Cumulative gains (%) in the course of 1 year for $\mathcal{VF}_1$

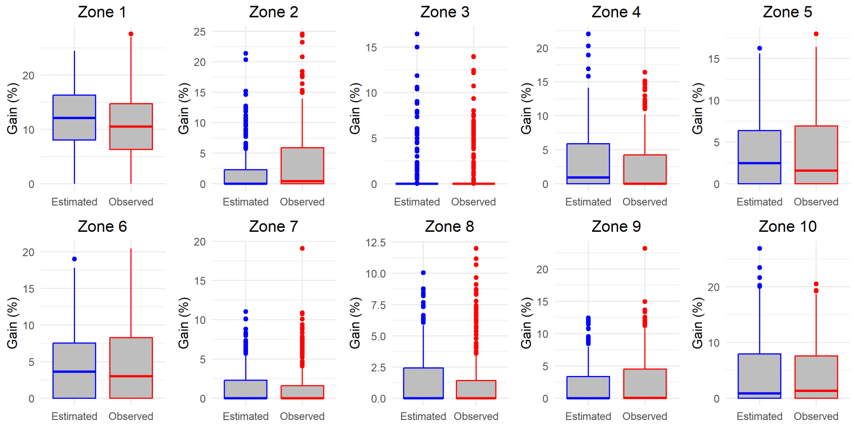
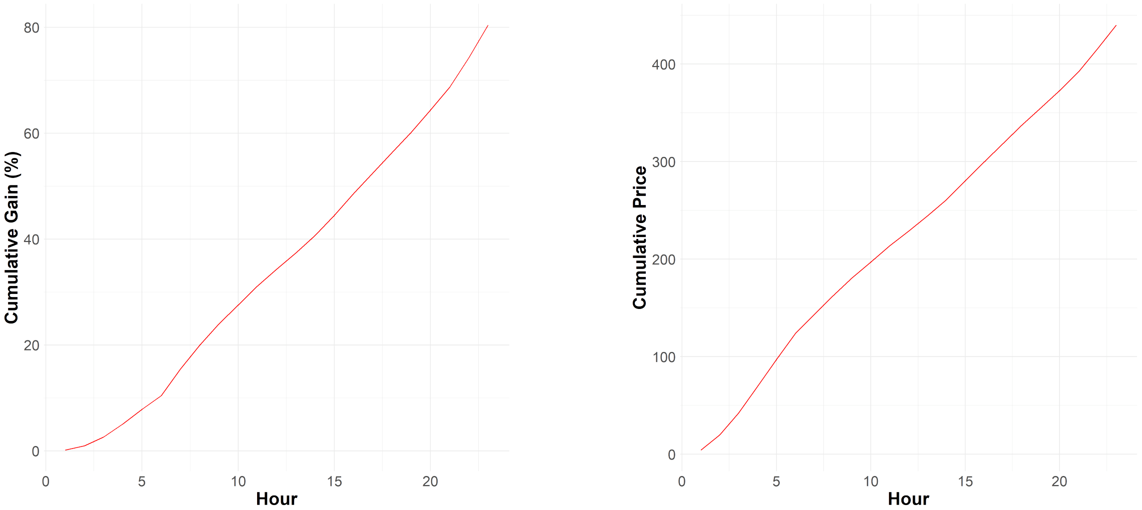


FIGURE 4.8: Boxplots of the gains (%) obtained in the course of 1 year for $\mathcal{VF}_1$

most proficient seller, extracting the highest revenue from data sales within the market. This observation aligns with our earlier correlation analysis, highlighting the significance of Zone 2 lags to the power forecasts of Zones 7, 8, and 9. Additionally, when considering each value function individually, the total sum of payments made by all buyers aligns precisely with the total sum of values acquired by all sellers utilizing the same value function. This equivalence is evident when comparing Table 4.6 to Table 4.7, underscoring the flow of financial transactions from buyers to sellers.



(A) Average cumulative gain (%) in a 24h resolution when using value function $\mathcal{VF}_1$

(B) Average cumulative price (%) in a 24h resolution when using value function $\mathcal{VF}_1$

FIGURE 4.9: Average cumulative gain and average cumulative price for 24-hour resolution forecasts

TABLE 4.7: Cumulative seller's gains in monetary units at the End of 1 Year

$\mathcal{S}$	$\mathcal{VF}_1$	$\mathcal{VF}_2$	$\mathcal{VF}_3$	$\mathcal{VF}_4$	$\mathcal{VF}_5$	$\mathcal{VF}_6$	**Mean**
Seller 1	4555	2548	4555	3940	483	1240	3179.8
Seller 2	5260	2856	5260	4526	757	1520	3896.5
Seller 3	5083	2776	5083	4342	448	1199	3698.5
Seller 4	4689	2435	4689	4064	392	1010	3588.2
Seller 5	4389	2099	4389	3758	469	987	3281.8
Seller 6	4714	2362	4714	4093	438	1137	3527.2
Seller 7	4833	2242	4833	4118	544	1062	3582.5
Seller 8	5089	2863	5089	4372	522	1352	3711.5
Seller 9	5131	2502	5131	4347	446	1163	3639.5
Seller 10	4648	2419	4648	3899	352	1226	3497.8
Sum	48391	25102	48391	41459	4851	11896	

Chapter 5

Conclusion

In conclusion, our research has explored the potential of collaborative data sharing among various Renewable Energy Source (RES) agents to significantly enhance RES forecasting accuracy across different time horizons, including short-term and day-ahead predictions. As established in the existing literature, the benefits of collaborative forecasting are clear. However, the implementation of such schemes hinges on creating economic incentives, particularly through data monetization, as RES agents, often competitors, may be reluctant to share their sensitive data without adequate benefits.

Existing data market solutions in the literature have fallen short in providing a comprehensive framework that can address the monetary requirements of both sellers and buyers while also safeguarding data privacy. In response to this gap, our proposed framework features an auction mechanism. This mechanism serves a dual purpose: it allows sellers to specify the minimum compensation they require for sharing their data, and it enables buyers to express their willingness to pay based on the forecast accuracy gain.

At the core of our framework is a market operator responsible for collecting data from all participating agents and generating forecasts. This operator employs a Splines Bid-Constrained Lasso Regression ($\mathcal{SBCLR}$) approach, aligning the forecasting mechanism with the mutually beneficial interests of buyers and sellers.

To evaluate our proposal, we conducted a case study focused on wind power generation 24 hours ahead forecasts for ten different wind farms in Australia. Our findings demonstrate that every agent can significantly enhance their forecasting skill by leveraging our collaborative framework. In a specific market configuration where all sellers bid at a unit price of 1,we showcased that within the span of a single day, an agent can achieve an accumulated gain of over 80% (i.e, a mean gain around 3.3% per hour), while incurring costs of less than 440 unit prices. Moreover, we highlighted the substantial monetary benefits for sellers, revealing that an agent can potentially extract 5260 price units under the same market configuration.

While our research shows the promise of our collaborative forecasting framework, there is still significant room for improvement. Future work in this field could involve the exploration of advanced techniques and strategies for buyers to refine their value functions, thereby maximizing their interests. Additionally, enhancing the bidding process from the seller′s perspective presents an avenue for further research. Sellers in our current framework base their bids solely on the utilization of their variables. A potential enhancement could involve allowing sellers to make bids according to the explanatory power of their variables, akin to the concept of value functions.

In summary, our research underscores the potential of our proposal to greatly improve forecasting skill for buyers within a collaborative framework that successfully addresses sellers′ interests. Through our innovative approach, we have bridged the gap between data monetization and collaborative forecasting, creating a win-win scenario for all participants in the RES market.

Appendix A

Appendix

A.1 Bid-Constrained Linear Regression

With the training data $\{y_{i,t}, z_t\}_{t=1}^{T}$ at hand, a logical approach to estimate the coefficients $\boldsymbol{\theta}_i = (\boldsymbol{\beta}^i, \boldsymbol{\eta}^i)$ is by addressing the subsequent sample-average approximation (SAA) problem:

$$\begin{aligned} &\arg_{\beta} \min \frac{1}{2T} \| y_{i,t} - \boldsymbol{z}_t \boldsymbol{\theta}_i \|_2^2 \\ \text{subject to } &\sum_{j=1}^{N} \left(\sum_{k=1}^{n_j} s_{j,k} \mathcal{I}\left(\beta_{j,k}^i\right) + \sum_{l=1}^{L} s_{j,l+n_j} \mathcal{I}\left(\eta_{l,j}^i\right) \right) \leq b_i, \end{aligned} \tag{A.1}$$

where $\boldsymbol{\theta}_i \in \mathcal{R}^p$ is the vector incorporating all elements in $\boldsymbol{\beta}^i$ and $\boldsymbol{\eta}^i$. The minimization problem highlighted above is termed the bid-constrained regression problem within the context of this article. When $\forall j, k, \quad s_{jk} = c \in \mathbb{R}$ this problem can be viewed as a generalized best subset selection problem [29]. Note that the budget constraint makes the problem (A.1) NP-hard. In fact, state-of-the-art algorithms for addressing the best subset problem encounter scalability issues with problems with more than 30 variables [30].

To tackle this issue, a two-step methodology that uses penalized regression techniques [31–33] can be employed. This approach unfolds in two distinct stages. Initially, these techniques are harnessed for model feature selection. In the subsequent stage, a search of the bid-constrained regression model is executed, using only the variables previously selected. Since many penalized regression techniques, such as LASSO regression in Section 2.2.2 preserve model consistency, this approach seems to yield promising outcomes as it saves a lot of computational resources. However, in many instances, the selected variables in the bid-constrained regression model satisfying the budget constraint may not be important for the regression task at hand (see the toy example in [26]). To solve the bid-constrained regression (A.1), we employ the same strategy used by [30] on a best subset selection problem by using projected gradient descent methods for the first-order convex optimization problems [34].

Let's consider the functions $g(\boldsymbol{\theta})$ and $\nabla g(\boldsymbol{\theta})$ as follows:

$$g(\boldsymbol{\theta}) = \frac{1}{2n}\|\boldsymbol{y_{i,t}} - \boldsymbol{Z\theta}\|_2^2$$

and

$$\nabla g(\boldsymbol{\theta}) = \left(\frac{\partial g(\boldsymbol{\theta})}{\partial \beta^i_{1,1}}, \ldots, \frac{\partial g(\boldsymbol{\theta})}{\partial \beta^i_{n_N,N}}, \frac{\partial g(\boldsymbol{\theta})}{\partial \eta^i_{1,1}}, \ldots, \frac{\partial g(\boldsymbol{\theta})}{\partial \eta^i_{N,L}}\right) = -\frac{1}{T}\boldsymbol{Z}^T(\boldsymbol{y}_{i,t} - \boldsymbol{Z\theta}),$$

with $\boldsymbol{\theta} \in \mathbb{R}^{p\times 1}$.

For any $\boldsymbol{\alpha}, \boldsymbol{\theta} \in \mathbb{R}^{p\times 1}$, and $L \geq \ell = \lambda_{\max}\left(\frac{\boldsymbol{Z}^T\boldsymbol{Z}}{n}\right)$, we can deduce:

$$\begin{aligned}\|\nabla g(\boldsymbol{\alpha}) - \nabla g(\boldsymbol{\theta})\|_2 &= \left\|\frac{1}{T}\boldsymbol{Z}^T\boldsymbol{Z}(\boldsymbol{\alpha} - \boldsymbol{\theta})\right\|_2 \\ &\leq \lambda \max\left(\frac{\boldsymbol{Z}^T\boldsymbol{Z}}{n}\right)\|\boldsymbol{\alpha} - \boldsymbol{\theta}\|_2 \leq L\|\boldsymbol{\alpha} - \boldsymbol{\theta}\|_2.\end{aligned} \tag{A.2}$$

Moreover, for any $\boldsymbol{\eta}, \boldsymbol{\theta} \in \mathbb{R}^{p\times 1}$, and $L \geq \ell = \lambda_{\max}\left(\frac{\boldsymbol{Z}^T\boldsymbol{Z}}{T}\right)$, let $Q_L(\boldsymbol{\eta}, \boldsymbol{\theta}) = g(\boldsymbol{\theta}) + \frac{L}{2}\|\boldsymbol{\eta} - \boldsymbol{\theta}\|_2^2 + \langle\nabla g(\boldsymbol{\theta}), \boldsymbol{\eta} - \boldsymbol{\theta}\rangle$. It can be observed that $Q_L(\boldsymbol{\eta}, \boldsymbol{\theta}) - g(\boldsymbol{\eta}) \geq \frac{L-\ell}{2}\|\boldsymbol{\theta} - \boldsymbol{\eta}\|_2^2 \geq 0$, leading to:

$$g(\boldsymbol{\eta}) \leq Q_L(\boldsymbol{\eta}, \boldsymbol{\theta}) = g(\boldsymbol{\theta}) + \frac{L}{2}\|\boldsymbol{\eta} - \boldsymbol{\theta}\|_2^2 + \langle\nabla g(\boldsymbol{\theta}), \boldsymbol{\eta} - \boldsymbol{\theta}\rangle, \tag{A.3}$$

which holds true for all $\boldsymbol{\theta}, \boldsymbol{\eta}$ and equality occurs at $\boldsymbol{\theta} = \boldsymbol{\eta}$. Consequently, given an approximate solution $\boldsymbol{\theta}^{(m)}$ to the problem (A.1), we can establish an upper bound for the function $g(\boldsymbol{\eta})$ using $Q_L\left(\boldsymbol{\eta}, \boldsymbol{\theta}^{(m)}\right)$. The solution can then be updated through:

$$\boldsymbol{\theta}^{(m+1)} \in \arg\min_{\eta} Q_L\left(\boldsymbol{\eta}, \boldsymbol{\theta}^{(m)}\right) \quad \text{subject to} \quad \sum_{j=1}^{N}\left(\sum_{k=1}^{n_j} s_{jk}w_{j+k} + \sum_{l=1}^{L} s_{j,l+n_j}w_{j+n_j+l}\right) \leq b_i,$$

which also represents a global minimizer of the subsequent problem:

$$\min_{\boldsymbol{\eta}}\left\|\boldsymbol{\eta} - \left(\boldsymbol{\theta}^{(m)} - \frac{1}{L}\nabla g\left(\boldsymbol{\theta}^{(m)}\right)\right)\right\|_2^2 \quad \text{subject to} \quad \sum_{j=1}^{N}\left(\sum_{k=1}^{n_j} s_{jk}w_{j+k} + \sum_{l=1}^{L} s_{j,l+n_j}w_{j+n_j+l}\right) \leq b_i. \tag{A.4}$$

As shown in [26], problem (A.4) is equivalent to a 0/1 knapsack problem as expressed by the following proposition.

Proposition A.1. *If $\hat{\boldsymbol{\theta}} = (\boldsymbol{\beta}^i, \boldsymbol{\eta}^i)$ is an optimal solution to the following problem:*

$$\begin{aligned} &\min_{\boldsymbol{\theta}} \|\boldsymbol{\theta} - a\|_2^2 \\ \textit{subject to} &\sum_{j=1}^{N} \left(\sum_{k=1}^{n_j} s_{j,k} \mathcal{I}\left(\boldsymbol{\beta}^i_{j,k}\right) + \sum_{l=1}^{L} s_{j,l+n_j} \mathcal{I}\left(\boldsymbol{\eta}^i_{l,j}\right) \right) \quad \leq b_i, \end{aligned} \tag{A.5}$$

then $\hat{\boldsymbol{\theta}} = a \circ \hat{\mathbf{W}}$ where $\circ$ denotes the entry-wise product of two vectors, and $\hat{\mathbf{W}} = \left(\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_p\right)$ with $p = \#\mathcal{Z}_t = \sum_{j=1}^{N} n_j + NL$ is the solution to the following 0-1 knapsack problem:

$$\begin{aligned} &\max_{w_1, w_2, \ldots, w_p} \sum_{j=1}^{p} a_j^2 w_j \\ \textit{subject to} &\sum_{j=1}^{N} \left(\sum_{k=1}^{n_j} s_{jk} w_{j+k} + \sum_{l=1}^{L} s_{j,l+n_j} w_{j+n_j+l} \right) \leq b_i \\ &w_1, w_2, \ldots, w_p \in \{0, 1\}. \end{aligned} \tag{A.6}$$

Proof. Since $\|\Theta - a\|^2 = \sum_j (\Theta_j - a_j)^2$ and $(\Theta_j - a_j)^2 \geq 0$, the maximization of $\|\Theta - a\|^2$ implies that if $a_j = 0$ the $\hat{\Theta}_j = 0$. Thus, the equation $\hat{\Theta}_j = a_j \hat{z}_j$ is satisfied if $a_j = 0$. Without loss of generality, let us assume that $a_j \neq 0$ for each j. Suppose $\hat{\Theta}$ is an optimal solution to the problem (A.1) and denote $\hat{S} = \{j \colon \hat{\Theta}_j \neq 0\}$. If $\hat{S} = \emptyset$, then $\hat{\Theta} = 0$. If $\hat{S} \neq \emptyset$, suppose that $\hat{\Theta}_j \neq a_j$ for some $j \in \hat{S}$, we can check that $\hat{b} = (\hat{\Theta}_1, \ldots, \hat{\Theta}_{j-1}, a_j, \hat{\Theta}_{j+1}, \ldots, \hat{\Theta}_p)$ is also a feasible solution and

$$\|\hat{\Theta} - a\|^2 = \|\hat{b} - a\|^2 + (\hat{\Theta}_j - a_j)^2 \geq \|\hat{b} - a\|^2. \tag{A.7}$$

Then we have a contradiction. Hence $\hat{\Theta}_j = a_j, \forall j \in \hat{S}$. Thus for optimal solution $\hat{\Theta}_j$, $\hat{\Theta}_j = 0$ or $\hat{\Theta}_j = a_j, \forall j$. Therefore, $\Theta_j = a \circ \hat{Z}$, where $\hat{Z}$ is the solution of the 0-1 knapsack problem in (A.6). □

The 0-1 knapsack problem has been deeply explored in the field of operations research (see, for instance, the book [35]). It entails selecting a subset of p items to maximize an objective function while ensuring that the combined weight remains under a predetermined capacity, which in this case is the buyer's bid b_i. Despite its status as an NP-hard problem, advancements in algorithms and hardware capabilities have facilitated efficient solutions. In this thesis, the dynamic programming algorithm, introduced in [36], is chosen to solve the 0-1 knapsack problem. Consider the problem (A.6). Let $D[j, w]$ represent the maximum attainable value for weights up to w using items of $\mathcal{Z}_t$ up to j. To lightness of notation purposes, we re-index the sellers bids: $s_i = \left(s_{1,1}, \ldots, s_{1,n_1+L}, \ldots, s_{N,1}, \ldots, s_{1,n_N+L}\right) =$

$\left(s_1, \ldots, s_j, \ldots, s_p\right)$. The recursive definition for $D[j, w]$ comes as follows:

$$\begin{aligned} &D[0, w] = 0 \\ &D[j, w] = D[j-1, w] \quad \text{if} \quad s_j > w \\ &D[j, w] = \max(D[j-1, w], D[j-1, w-c_j] + a_j^2) \quad \text{if} \quad s_j \leq w \end{aligned} \tag{A.8}$$

The solution can be found by evaluating $D[p, C]$. Dynamic programming algorithm works only with non-negative integers weights, i.e, $s_j, b_i \geq 0$, If there is some s_j or b_i values non-integers, we can employ a scaling technique that uniformly multiplies both the non-integer costs and the budget by a constant factor, thereby converting them into integers.

A.2 Splines Bid-Constrained Regression

Using the notation in Chapter 3, each variable $z \in \mathcal{Z}_t$ is transformed into a group of M variables through a spline transformer, where $M = D + K + 1$ is the sum of the spline order D with the number of knots K plus 1. After the spline transformer, the original set:

$$\mathcal{Z}_t = \left\{ \underbrace{X_{1,t}^1, \ldots, X_{1,t}^{n_1}, Y_{1,t_0-1}, \ldots, Y_{1,t_0-L}}_{\text{power plant 1}}, \ldots, \underbrace{X_{N,t}^1, \ldots, X_{N,t}^{n_N}, Y_{N,t_0-1}, \ldots, Y_{N,t_0-L}}_{\text{power plant N}} \right\}. \tag{A.9}$$

is transformed into

$$\tilde{\mathcal{Z}}_t = \left\{ \underbrace{X_{1,1,t}^1, \ldots, X_{1,M,t}^1}_{\text{group related to } X_{1,t}^1}, \ldots, \underbrace{X_{1,1,t}^{n_1}, \ldots, X_{1,M,t}^{n_1}}_{\text{group related to } X_{1,t}^{n_1}}, \ldots \right\} \tag{A.10}$$

that is constituted by p groups of variables, where $p = \sum_{j=1}^{N} n_j + NL$ is the number of variables in $\mathcal{Z}_t$.

We consider seller j wants to receive $s_{j,k}$ if its k^{th} variable is used. Since we apply splines, we consider that if at least one of the variables in the group related to the k^{th} variable is used, then seller j will receive $s_{j,k}$. Therefore, the payment of buyer i, p_i, needs to cover the cost of the model,

$$\hat{\boldsymbol{\beta}}_i,\hat{\boldsymbol{\eta}}_i = \underset{\boldsymbol{\beta},\boldsymbol{\eta}}{\operatorname{argmin}} \frac{1}{2}\sum_{t=1}^{T}\left(\boldsymbol{y}_{i,t} - f\left(\tilde{\mathbf{z}}_t;\boldsymbol{\beta}_i,\boldsymbol{\eta}_i\right)\right)^2 + \lambda\sum_{m=1}^{M}\left(\sum_{j=1}^{N}\left[\sum_{k=1}^{n_j}|\beta^i_{k,j,m}| + \sum_{l=1}^{L}|\eta_{l,j,m}|\right]\right)$$
$$\text{s.t.} \sum_{j=1}^{N}\left(\sum_{k=1}^{n_j} s_{jk}\left[1-\prod_{m=1}^{M}\left(1-\mathcal{I}\left(\beta^i_{j,k,m}\right)\right)\right] + \sum_{l=1}^{L} s_{j,l+n_j}\left[1-\prod_{m=1}^{M}\left(1-\mathcal{I}\left(\eta^i_{l,j,m}\right)\right)\right]\right) \leq p_i. \tag{A.11}$$

Following the reasoning in (A.2)– (A.4), the problem (A.11) is re-written as:

$$\begin{aligned} \min_{\boldsymbol{\Theta}_i} \frac{1}{2}\|\boldsymbol{\Theta}_i - \boldsymbol{a}\|_2^2 + \lambda\sum_{m=1}^{M}\left(\sum_{j=1}^{N}\left[\sum_{k=1}^{n_j}|\beta^i_{k,j,m}| + \sum_{l=1}^{L}|\eta_{l,j,m}|\right]\right) \\ \text{subject to} \quad \sum_{j=1}^{N}\left(\sum_{k=1}^{n_j} s_{jk}\left[1-\prod_{m=1}^{M}\left(1-\mathcal{I}\left(\beta^i_{j,k,m}\right)\right)\right] + \right. \\ \left. \sum_{l=1}^{L} s_{j,l+n_j}\left[1-\prod_{m=1}^{M}\left(1-\mathcal{I}\left(\eta^i_{l,j,m}\right)\right)\right]\right) \leq p_i. \end{aligned} \tag{A.12}$$

Then, as shown in [26], problem (A.12) is also equivalent to a 0/1 knapsack problem as expressed by the following proposition.

Proposition A.2. *If $\hat{\boldsymbol{\Theta}}_i = (\hat{\boldsymbol{\beta}}^i, \hat{\boldsymbol{\eta}}^i)$ is an optimal solution to (A.12), then*

$$\hat{\boldsymbol{\Theta}} = sign(\boldsymbol{a} - \lambda) \circ (|\boldsymbol{a}| - \lambda)_+ \circ \hat{\boldsymbol{Z}},$$

where $\hat{\boldsymbol{Z}} = \left(\hat{z}_{1,1}\mathbf{1}_M, \hat{z}_{1,2}\mathbf{1}_M, \ldots, \hat{z}_{jk}\mathbf{1}_M\right)^T$, $\mathbf{1}_M$ is the row vector of M 1's, and $\hat{z}_{1,1}, \hat{z}_{1,2}, \ldots, \hat{z}_{j,k}$ is the solution to the following 0-1 knapsack problem:

$$\max_{z_{1,1},z_{1,2},\ldots,z_{jk}} \sum_{j=1}^{N}\sum_{k=1}^{n_j} \underbrace{\left(\sum_{m=1}^{M} \frac{a^2_{j,k,m} - 2\lambda|a_{j,k,m}| + \lambda^2}{2} \cdot \frac{1 + sign(|a_{j,k,m}| - \lambda)}{2}\right)}_{\mu_{j,k}} z_{j,k} \tag{A.13}$$
$$\text{subject to } \left(z_{1,1}, \ldots, z_{j,k}, \ldots, z_{N,L+n_L}\right) \cdot \left(s_{1,1}, \ldots, s_{j,k}, \ldots, s_{N,L+n_N}\right) \leq p_i.$$

Bibliography

[1] Eurostat, "Renewable energy statistics," Retrieved from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Renewable_energy_statistics, 2021. [Cited on page 1.]

[2] T. Gneiting, K. Larson, K. Westrick, M. G. Genton, and E. Aldrich, "Calibrated probabilistic forecasting at the stateline wind energy center," *Journal of the American Statistical Association*, vol. 101, no. 475, pp. 968–979, 2006. [Online]. Available: https://doi.org/10.1198/016214506000000456 [Cited on page 1.]

[3] Q. Zhu, J. Chen, D. Shi, L. Zhu, X. Bai, X. Duan, and Y. Liu, "Learning temporal and spatial correlations jointly: A unified framework for wind speed prediction," *IEEE Transactions on Sustainable Energy*, vol. 11, no. 1, pp. 509–523, 2019. [Cited on page 1.]

[4] L. Cavalcante, R. J. Bessa, M. Reis, and J. Browell, "Lasso vector autoregression structures for very short-term wind power forecasting," *Wind Energy*, vol. 20, no. 4, pp. 657–675, 2017. [Cited on page 1.]

[5] L. Cavalcante and R. J. Bessa, "Solar power forecasting with sparse vector autoregression structures," in *2017 IEEE Manchester PowerTech*. IEEE, 2017, pp. 1–6. [Cited on page 1.]

[6] C. Goncalves, R. J. Bessa, and P. Pinson, "Privacy-preserving distributed learning for renewable energy forecasting," *IEEE Transactions on Sustainable Energy*, vol. 12, no. 3, pp. 1777–1787, 2021. [Cited on pages 2 and 37.]

[7] J. Parra-Arnau, "Optimized, direct sale of privacy in personal data marketplaces," *Information Sciences*, vol. 424, pp. 354–384, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025517310022 [Cited on page 2.]

[8] S. Mehta, M. Dawande, G. Janakiraman, and V. Mookerjee, "How to sell a dataset? pricing policies for data monetization," *Information Systems Research*, pp. 679–679, 2019. [Online]. Available: https://ssrn.com/abstract=3333296 [Cited on page 2.]

[9] A. Agarwal, M. Dahleh, and T. Sarkar, "A marketplace for data: An algorithmic solution," 2019. [Cited on page 2.]

[10] C. Gonçalves, P. Pinson, and R. J. Bessa, "Towards data markets in renewable energy forecasting," *IEEE Transactions on Sustainable Energy*, vol. 12, no. 1, pp. 533–542, 2021. [Cited on pages 2 and 17.]

[11] W.-Y. Chang *et al.*, "A literature review of wind forecasting methods," *Journal of Power and Energy Engineering*, vol. 2, no. 04, p. 161, 2014. [Cited on page 8.]

[12] C. S. d. S. Gonçalves, "Renewable energy forecasting – extreme quantiles, data privacy and monetization," Ph.D. dissertation, University of Porto, 2021. [Cited on page 8.]

[13] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman, "Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond," pp. 896–913, 2016. [Cited on page 8.]

[14] M. Milligan, M. Schwartz, and Y.-h. Wan, "Statistical wind power forecasting models: Results for us wind farms," National Renewable Energy Lab.(NREL), Golden, CO (United States), Tech. Rep., 2003. [Cited on page 9.]

[15] M. S. Miranda and R. W. Dunn, "One-hour-ahead wind speed prediction using a bayesian methodology," in *2006 IEEE Power Engineering Society General Meeting*. IEEE, 2006, pp. 6–pp. [Cited on page 9.]

[16] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," *Advances in neural information processing systems*, vol. 25, 2012. [Cited on page 10.]

[17] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2009. [Cited on page 13.]

[18] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003. [Cited on page 16.]

[19] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," *Data classification: Algorithms and applications*, p. 37, 2014. [Cited on page 16.]

[20] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*. Ieee, 2015, pp. 1200–1205. [Cited on page 17.]

[21] X. Gu, J. Guo, L. Xiao, and C. Li, "Conditional mutual information-based feature selection algorithm for maximal relevance minimal redundancy," *Applied Intelligence*, vol. 52, no. 2, pp. 1436–1447, 2022. [Cited on page 17.]

[22] L. S. Shapley *et al.*, "A value for n-person games," 1953. [Cited on page 18.]

[23] P. Pinson, L. Han, and J. Kazempour, "Regression markets and application to energy forecasting," *Top*, vol. 30, no. 3, pp. 533–573, 2022. [Cited on pages 19 and 20.]

[24] L. Han, P. Pinson, and J. Kazempour, "Trading data for wind power forecasting: A regression market with lasso regularization," *Electric Power Systems Research*, vol. 212, p. 108442, 2022. [Cited on page 21.]

[25] P. Pinson, L. Han, and J. Kazempour, "Regression markets and application to energy forecasting," *Top*, vol. 30, no. 3, pp. 533–573, 2022. [Cited on page 21.]

[26] G. Yu, H. Fu, and Y. Liu, "High-dimensional cost-constrained regression via nonconvex optimization," *Technometrics*, vol. 64, no. 1, pp. 52–64, 2022. [Cited on pages 30, 32, 55, 56, and 59.]

[27] F. X. Diebold and R. S. Mariano, "Comparing predictive accuracy," *Journal of Business & economic statistics*, vol. 20, no. 1, pp. 134–144, 2002. [Cited on page 45.]

[28] D. Harvey, S. Leybourne, and P. Newbold, "Testing the equality of prediction mean squared errors," *International Journal of Forecasting*, vol. 13, no. 2, pp. 281–291, 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169207096007194 [Cited on page 46.]

[29] A. Miller, *Subset selection in regression*. CRC Press, 2002. [Cited on page 55.]

[30] D. Bertsimas, A. King, and R. Mazumder, "Best subset selection via a modern optimization lens," *The Annals of Statistics*, vol. 44, no. 2, pp. 813 – 852, 2016. [Online]. Available: https://doi.org/10.1214/15-AOS1388 [Cited on pages 55 and 56.]

[31] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006. [Online]. Available: https://doi.org/10.1198/016214506000000735 [Cited on page 55.]

[32] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996. [Online]. Available: http://www.jstor.org/stable/2346178

[33] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, vol. 38, no. 2, pp. 894 – 942, 2010. [Online]. Available: https://doi.org/10.1214/09-AOS729 [Cited on page 55.]

[34] Y. Nesterov, "Gradient methods for minimizing composite functions," *Mathematical programming*, vol. 140, no. 1, pp. 125–161, 2013. [Cited on page 56.]

[35] S. Martello and P. Toth, *Knapsack problems: algorithms and computer implementations*. John Wiley & Sons, Inc., 1990. [Cited on page 57.]

[36] R. Bellman, "Dynamic programming," *Science*, vol. 153, no. 3731, pp. 34–37, 1966. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.153.3731.34 [Cited on page 57.]