This paper appears in: IEEE Transactions on Sustainable Energy Print ISSN: 1949-3029 Online ISSN: 1949-3037 Digital Object Identifier: 10.1109/TSTE.2020.3009615

Towards Data Markets in Renewable Energy Forecasting

Carla Gonçalves, Pierre Pinson, *Fellow, IEEE*, and Ricardo J. Bessa, *Senior Member, IEEE*

***Abstract*—Geographically distributed wind turbines, photovoltaic panels and sensors (e.g., pyranometers) produce large volumes of data that can be used to improve renewable energy sources (RES) forecasting skill. However, data owners may be unwilling to share their data, even if privacy is ensured, due to a form of prisoner's dilemma: all could benefit from data sharing, but in practice no one is willing to do do. Our proposal hence consists of a data marketplace, to incentivize collaboration between different data owners through the monetization of data. We adapt here an existing auction mechanism to the case of RES forecasting data. It accommodates the temporal nature of the data, i.e., lagged time-series act as covariates and models are updated continuously using a sliding window. A test case with wind energy data is presented to illustrate and assess the effectiveness of such data markets. All agents (or data owners) are shown to benefit in terms of higher revenue resulting from the combination of electricity and data markets. The results support the idea that data markets can be a viable solution to promote data exchange between RES agents and contribute to reducing system imbalance costs.**

***Index Terms*—Collaborative forecasting, data marketplace, data pricing, renewable energy, electricity market.**

Notation	Description
ρ	Electricity profit function
π_t^s	Spot price
$\pi_t^{\uparrow}$, $\pi_t^{\downarrow}$	Imbalance price for upward / downward regulation
$\lambda_t^{\uparrow}$, $\lambda_t^{\downarrow}$	Regulation unit cost for upward / downward directions
$C_t^{\uparrow/\downarrow}$	Imbalance cost
α_t^*	Nominal level which minimizes $C_t^{\uparrow/\downarrow}$
$\hat{F}_{i,t}^{-1}(\alpha_t^*)$	Forecasted conditional quantile for nominal level α_t^*
$\hat{\psi}_t^{\uparrow}$, $\hat{\psi}_t^{\downarrow}$	Forecasted upward / downward regulation price
$\hat{p}_t^{\uparrow}$, $\hat{p}_t^{\downarrow}$	Probability of up/downward regulation at time t
N	Number of RES power agents
$\mathcal{A}$	Overall set of power plants, $\mathcal{A}=\{1,\ldots,N\}$
T	Number of historical records
H	Length of the time horizon
$x_{i,t}$	Power measurements for RES agent i at time t
$\hat{q}_{\alpha_t^*}^i$	Forecasted quantile α_t^* for site $i \in \mathcal{A}$ at time t
β_-^-	Linear quantile regression coefficients
$\mathbf{x}_i^{\mathrm{S}}$ (or $\mathbf{x}_i^{\mathrm{B}}$)	Data from seller (or buyer) i
$\mathbf{X}^{\mathrm{S}}$	Data from all sellers, $\mathbf{X}^{\mathrm{S}}=[\mathbf{x}_1^{\mathrm{S}},\ldots,,\mathbf{x}_N^{\mathrm{S}}] \in \mathbb{R}^{T\times N}$
$\mathcal{M}_i$	Forecasting model for power production of agent i
$\mathcal{G}_i$	Gain function for buyer i
μ_i	Private valuation for each unit gain
b_i	Public bid price (buyer i is willing to pay $b_i \leq \mu_i$)
p_i	Data market price for buyer i
$\mathcal{U}_i$	Value (or utility) function for buyer i
$\mathcal{PF}$	Market price update function (price for the buyer)
$\mathcal{RF}$	Revenue function (price to be paid by buyers)
$\mathcal{AF}$	Allocation function (variables allocation given b_i, p_i)
$\mathcal{PD}$	Payment division function (division by sellers)
$\mathcal{N}(0,\sigma^2)$	Normal distribution with σ standard deviation
$\mathcal{SM}$	Similarity function (similarity between two vectors)
$p_{\min}, p_{\max}$	Minimum and maximum possible data market prices
Δ_p	Increments on possible data market prices
$\mathcal{B}_p$	All possible market prices
$\psi_i(m)$	Fraction of money paid by buyer i allocated to agent m
Δ	Length of the period used to estimate the gain
K	Number of repetitions in the Shapley Approximation

C. Gonçalves is with INESC Technology and Science (INESC TEC), Campus da FEUP, 4200-465 Porto, Portugal and Faculty of Sciences of the University of Porto (FCUP), Portugal (e-mail: carla.s.goncalves@inesctec.pt). P. Pinson is with the Technical University of Denmark, 2800 Kongens Lyngby, Denmark (e-mail: ppin@elektro.dtu.dk). R.J. Bessa is with INESC TEC, Campus da FEUP, 4200-465 Porto, Portugal (e-mail: ricardo.j.bessa@inesctec.pt).

The research leading to this work is being carried out as a part of the Smart4RES project (European Union's Horizon 2020, No. 864337). The sole responsibility of this publication lies with the author. The European Union is not responsible for any use that may be made of the information contained therein. C. Gonçalves was supported by the Portuguese funding agency, FCT (Fundação para a Ciência e a Tecnologia), within the Ph.D. grant PD/BD/128189/2016 with financing from POCH (Operational Program of Human Capital) and the EU.

I. Introduction

A LARGE amount of data is being collected from geographically distributed renewable energy sources (RES) such as wind turbines and photovoltaic (PV) panels. These data include power generation and weather measurements like air temperature, wind speed and direction, irradiation, etc.

Recent literature suggests that time-series data from spatially distributed RES agents can improve forecasting skill for different time horizons. For instance, a spatial grid of numerical weather predictions (NWP) can improve days-ahead forecasts [1]; turbine-level data can improve the day-ahead forecasting skill of wind energy through density forecasts generated for all wind turbines with spatial dependency structure modelled via copula theory [2]. Geographically distributed time-series data can improve forecasting skill up to 6 hours-ahead for wind [3] and solar energy [4]. In fact, hours-ahead forecasts will become a crucial input for decision-aid as intraday electricity markets (e.g., European cross-border intraday – XBID) become increasingly important for RES technology.

However, since RES agents are most likely competitors in the same electricity market, they are unwilling to share data, particularly power measurements, even if data privacy is ensured. An effective way to encourage agents to share their data is through monetary compensation [5], [6]. A "secondary" market to trade data is necessary to monetize RES forecasting data. Moreover, this data market should operate in a way that, after some iterations, agents realize which data is relevant to improve its gain, so that sellers are paid according to their data. The buyers' gain should be a function of the forecast

accuracy and value in a specific use case, e.g. imbalance costs reduction in electricity market bidding. It is important to mention that a RES plant owner can buy, from a vendor, NWP for neighbor power plants, but not their power measurements (or forecasts) that contain relevant information to improve hours-ahead forecasting skill. By joining a data market, the data owner can also sell this additional data (e.g., NWP for nearby sites) and decrease its purchasing cost. Moreover, there are no guarantees that NWP for other locations are cheaper than buying information from a data market where the payment is a function of the forecasting skill improvement. In fact, when buying NWP from vendors, there are no *a priori* guarantees of improvement in the existing forecasting model.

A data auction mechanism is proposed in [7] where sellers compute the privacy cost of selling the data and then send it to a buyer that computes a utility score associated with the data. Several iterations are performed until a Bayesian Nash equilibrium is reached. A market mechanism is introduced in [8] to solve a social welfare maximization problem that defines the data allocation and corresponding price. In this case, data are only shared after payment. However, in order to compute data price, a utility function, which depends solely on quantity (i.e., data quality is not considered), is assumed to exist. This is not directly applicable to time-series forecasting with RES spatial data where correlated data from neighbor agents might be less informative than data from more distant agents (or sites). Furthermore, in [9], the impact of a strong correlation between data of different agents is analyzed as a negative externality from data sharing, e.g., buying the data from user A may reveal too much information about user B and the market price tends to zero (i.e., no value for data privacy). Different policies (e.g., "de-correlation") and regulatory schemes to data markets are proposed and analyzed. In [10], evolutionary game theory is combined with blockchain smart contracts to dynamically adjust incentives and participation costs in data sharing. In the energy domain, a market is proposed in [11] for smart meter data. The proposed game theory mechanism works as follows: (i) the consumer maximizes its reward from sharing consumption data; (ii) data aggregator expects to receive more money from the data analyst, rather than providing incentives to consumers; (iii) data analyst is interested in high quality data at the lowest possible cost. Also for smart meter data, a blockchain smart contract is designed in [12] to define a set of rules for data access control and reward against privacy risk. In both works, the payment is directly related to the privacy loss and not directly linked to the gain obtained from using this data in a specific decision-making problem. The concept of pricing data as a function of privacy loss is further discussed in [13], where the impact of sellers' risk attitude is analyzed.

Moreover, the temporal nature of RES forecasting also needs to be considered. An auction mechanism for time-series data is proposed in [14] where privacy is guaranteed with data distortion by adding random noise, in a way that preserves some time-series statistics and avoids the original series to be recreated when sold incrementally. Buyers ask for specific features together with the maximum noise they are willing to tolerate. Based on the level of noise, the market operator determines the privacy loss for selected data owners and sets the market prices to compensate them for the privacy loss. Buyer gain is not considered.

Since RES agents may be unwilling to share their data with competitors and mask of sensible data through noise addition involves a trade-off between privacy and accuracy [15], the framework from [16] offers an appealing alternative based on cooperative game theory. As far as we know, this is the first work to consider a marketplace where data owners purchase forecasts and pay according to resulting forecasting accuracy. This avoids the confidentiality problem of sharing raw data directly. Cooperation between sellers is done through a market operator who receives all agents data and prepares forecasts: (i) sellers with similar information receive similar revenue, (ii) the market price is a function of the buyer's benefit, and so the buyer does not pay if there is no improvement in the forecasting skill, (iii) buyers pay according to incremental gain, and (iv) buyers purchase forecasts, instead of features, and have no knowledge about which datasets were used to produce these forecasts. Sellers' loss is assumed to be zero.

Nevertheless, adaptions are necessary since time-series models require temporal updates of the input variables. Thus, the present paper presents the following original contributions:

i) The approach from [16] is extended for a sliding window environment and the gain function is adapted for RES forecasting and bidding in the electricity market.
ii) With geographically distributed time-series data, buyers want to integrate private and local data into the market operator's forecasts in order to avoid paying for highly-correlated data from close neighbors and this requirement is covered in the proposed approach – the approach in [16] does not consider RES agents with internal forecasting models and for which highly-correlated features might provide no improvement.
iii) Agents trade between themselves, i.e. sellers are buyers and buyers are sellers – sellers and buyers are independent agents in [16], thus adaptions are required to ensure that agents do not pay for their own or redundant data.

To the best of our knowledge, this is the first work to describe an algorithmic solution for data markets that enable different RES agents to sell data (historical power production, NWP, etc.) and buy forecasts of their power production, and where the economic value of this data is fundamentally related to imbalance cost reduction in electricity markets.

The paper is organized as follows. Section II formalizes the electricity market and forecasting framework. Section III proposes a data market for RES forecasting. Then, three test cases are considered in Section IV, two with synthetic data and another with Nordpool wind energy data. The work concludes in Section V.

II. Electricity Market and Forecast Framework

RES market agents aim to minimize imbalance costs (i.e., maximize electricity market profit) by improving forecasting skill. This section presents the market profit function and the formulation of the forecasting problem.

A. Electricity Market Profit Function

In a typical electricity market with dual price imbalance settlement [17], the profit function of a RES market agent, with power measurement x_t and forecast $\hat{x}_t$, is determined for each time step t as

$$\rho(\hat{x}_t, x_t) = \pi_t^s x_t - C_t^{\uparrow/\downarrow}, \tag{1}$$

where

$$C_t^{\uparrow/\downarrow} = \begin{cases} \lambda^{\uparrow}(\hat{x}_t - x_t), & \hat{x}_t > x_t \\ -\lambda^{\downarrow}(\hat{x}_t - x_t), & \hat{x}_t < x_t, \end{cases} \tag{2}$$

$$\lambda_t^{\uparrow} = \max(0, \pi_t^{\uparrow} - \pi_t^s), \tag{3}$$

$$\lambda_t^{\downarrow} = \max(0, \pi_t^s - \pi_t^{\downarrow}), \tag{4}$$

with π_t^s, $\pi_t^{\uparrow}$ and $\pi_t^{\downarrow}$ denoting the spot price, imbalance price for upward and downward regulation, respectively; $\lambda_t^{\uparrow}$ and $\lambda_t^{\downarrow}$ give the regulation unit cost for upward and downward directions.

For simplicity, generation costs are not considered in the profit function ρ. Furthermore, by calculating the derivative of the expected regulation cost with respect to the bid [17], it is possible to conclude that forecasts that maximize the profit in (1) do not correspond to the expected value of x_t, instead, they correspond to the quantile of the following nominal level,

$$\alpha_t^* = \frac{\hat{\lambda}_t^{\downarrow}}{\hat{\lambda}_t^{\uparrow} + \hat{\lambda}_t^{\downarrow}}, \tag{5}$$

where $\hat{\lambda}_t^{\uparrow}, \hat{\lambda}_t^{\downarrow}$ are deterministic forecasts for $\lambda_t^{\uparrow}, \lambda_t^{\downarrow}$.

This means that the optimal bid (i.e., the one that minimizes the expected imbalance costs in (2)) for a RES agent $i{=}1, \ldots, N$ is given by $\hat{F}_{i,t}^{-1}(\alpha_t^*)$ [17], where $\hat{F}_{i,t}^{-1}$ is the inverse of the forecasted cumulative distribution function or, in other words, corresponds to the forecasted conditional quantile for nominal level α_t^*. These analytical formulas for optimal bidding can be generalized for other situations, such as and joint offer of energy and reserve capacity [18].

In order to compute the "optimal" quantile from (5), a forecast of the regulation unit costs is required. Since we do not aim to propose a new forecasting model for imbalance prices, the Holt-Winters model described in [19] was used in this work. The upward regulation unit cost is estimated as the product between the forecasted upward regulation price ($\hat{\psi}_t^{\uparrow}$) and the probability of the system to be in upward regulation direction ($\hat{p}_t^{\uparrow}$), i.e.

$$\hat{\lambda}_t^{\uparrow} = \hat{\psi}_t^{\uparrow} \hat{p}_t^{\uparrow}. \tag{6}$$

Similarly,

$$\hat{\lambda}_t^{\downarrow} = \hat{\psi}_t^{\downarrow} \hat{p}_t^{\downarrow}, \tag{7}$$

where $\hat{p}_t^{\downarrow} = 1 - \hat{p}_t^{\uparrow}$ since we only care about relative probabilities for upward and downward regulation. The regulation prices are forecasted by

$$\hat{\psi}_{t|t-1}^i = \begin{cases} \eta \hat{\psi}_{t-1|t-2}^i + (1-\eta)(\lambda_{t-1}^i - \hat{\psi}_{t-1|t-2}^i), & |\lambda_{t-1}^i| > 0 \\ \hat{\lambda}_{t-1|t-2}^i, & |\lambda_{t-1}^i| = 0, \end{cases} \tag{8}$$

for $i \in \{\uparrow, \downarrow\}$, and the probability of system regulation direction by

$$\hat{p}_{t|t-1}^{\uparrow} = \begin{cases} \eta \hat{p}_{t-1|t-2}^{\uparrow} + (1-\eta)(p_{t-1}^{\uparrow} - \hat{p}_{t-1|t-2}^{\uparrow}), & p_{t-1}^{\uparrow} \neq 0.5 \\ \hat{p}_{t-1|t-2}^{\uparrow}, & p_{t-1}^{\uparrow} = 0.5, \end{cases} \tag{9}$$

where $\eta \in [0, 1[$ is a smoothing factor, and

$$p_{t-1}^{\uparrow} = \begin{cases} 1, & \lambda_{t-1}^{\uparrow} > \lambda_{t-1}^{\downarrow} \\ 0.5, & \lambda_{t-1}^{\uparrow} = \lambda_{t-1}^{\downarrow} \\ 0, & \lambda_{t-1}^{\uparrow} < \lambda_{t-1}^{\downarrow}. \end{cases} \tag{10}$$

Initialization of $p_0^{\uparrow}$, $\lambda_0^{\uparrow}$ and $\lambda_0^{\downarrow}$ is required, and η is estimated by minimizing the mean of squared residuals.

Given the forecasted values for regulation unit costs, the last step is to forecast the quantile with nominal level α_t^* using linear quantile regression as described in the next subsection. Note that here we are assuming a price-taker RES agent for the regulation market.

B. Formulation of the RES Forecasting Problem

In this work, we formulate a very short-term forecasting problem (up to 6h-ahead) involving multiple RES power plants. The forecasting model only uses recent measurements at all sites of interest, but longer time horizons with extra variables, such as grid of NWP [1] and turbine-level data [2], may also be considered using the same framework.

Assume that RES power plants generation data are collected at N sites, and $x_{i,t}$ denotes the power measurement at site i and time t, $i \in \mathcal{A}$, $t = 1, \ldots, T$, where T is the number of time steps in the dataset and $\mathcal{A}{=}\{1, \ldots, N\}$ is the overall set of power plants. We consider that these agents operate a single power plant, but the case where agents operate a portfolio of RES power plants may also be elaborated using the same framework.

The linear quantile regression model is a standard and straightforward method of conditional quantile estimation [20]. For very short-term forecasts, satisfactory results may be obtained by using the L most recent observations, as shown in [4] and [3] for both solar and wind energy.

In this case, the quantile α_{t+h}^* of power $x_{i,t+h}$ in site $i \in \mathcal{A}$ is expressed as

$$\hat{q}_{\alpha_{t+h}^*}^i = \beta_{0,i}^{(\alpha_{t+h}^*)} + \sum_{\ell=1}^{L} \Big(\underbrace{\sum_{j \in \mathcal{A} \setminus \{i\}} \hat{\beta}_{j,i,\ell}^{(\alpha_{t+h}^*)} x_{j,t-\ell}}_{\text{data from the market}} + \underbrace{\hat{\beta}_{i,i,\ell}^{(\alpha_{t+h}^*)} x_{i,t-\ell}}_{\text{own data}} \Big), \tag{11}$$

where $h \leq 6$ is the forecasting horizon, $\beta_{0,i}^{(\alpha_t^*)}$, $\beta_{j,i,\ell}^{(\alpha_t^*)}$ and $\beta_{i,i,\ell}^{(\alpha_t^*)}$ are the unknown coefficients, estimated through the minimization of the pinball loss function [20].

III. Data Market for RES Forecasting

This section proposes a no-regret auction mechanism for trading RES forecasts, as illustrated in Fig. 1. The buyers should never buy data because its value is unknown before

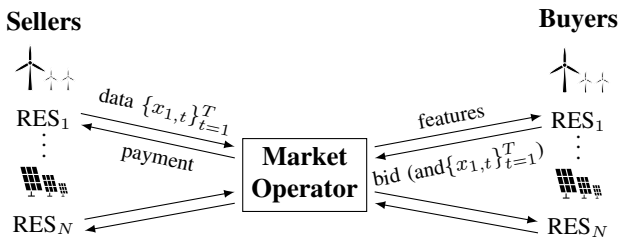


Fig. 1. Proposed data market framework.

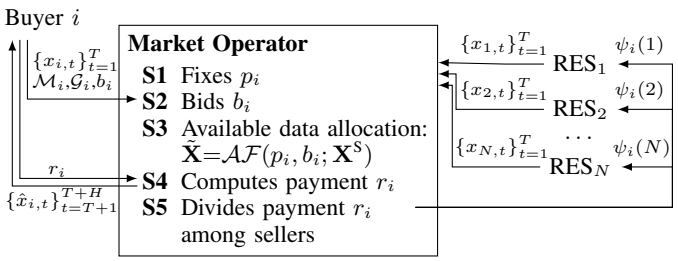


Fig. 2. Data market mechanism at time $t = T$.

using it for a forecasting task. Instead, they should purchase forecasts of their power production and pay according to the obtained forecasting accuracy. The data market formulation is inspired by the cooperative game in [16] and described in the following subsection in order to be self-content.

In addition to large RES power plants, this data market is also open to prosumers. Interestingly, data traders can also be interpreted as *data prosumers*, i.e. data owners that consume and supply data.

A. Data Market Agents

Like any standard market, the data market has three types of agents described in this subsection: sellers, buyers and market operator.

1) Sellers: A seller i observes and sells sample $\mathbf{x}_i^{\mathrm{S}}=\{x_{i,t}\}_{t=1}^{T}$, $\mathbf{x}_i^{\mathrm{S}} \in \mathbb{R}^T$, $i=1,\ldots,N$. Additionally, sellers have no idea of the forecasting methods that will use their data and simply aim to maximize their revenue. The set of features provided by all sellers is denoted by $\mathbf{X}^{\mathrm{S}}=[\mathbf{x}_1^{\mathrm{S}},\ldots,\mathbf{x}_N^{\mathrm{S}}]$, $\mathbf{X}^{\mathrm{S}} \in \mathbb{R}^{T\times N}$.

2) Buyers: A buyer i observes and seeks to improve sample $\mathbf{x}_i^{\mathrm{B}} = \{x_{i,t}\}_{t=1}^{T}$, $i = 1,\ldots,N$, and enters the data market to purchase the collection of features that allow a certain gain when forecasting $\{x_{i,t}\}_{t=T+1}^{T+H}$, through a selected method (statistical model) $\mathcal{M}_i$, $H \geq 1$. Buyers naturally have a local forecasting model $\mathcal{M}_i(\mathbf{x}_i^{\mathrm{B}})$, and enter the market to improve it with more features from the other agents, $\mathbf{X}_{\neg i}^{\mathrm{S}}$, where $\mathbf{X}_{\neg i}^{\mathrm{S}}=[\mathbf{x}_1^{\mathrm{S}},\ldots,\mathbf{x}_{i-1}^{\mathrm{S}},\mathbf{x}_{i+1}^{\mathrm{S}},\ldots,\mathbf{x}_N^{\mathrm{S}}]$, $\mathbf{X}_{\neg i}^{\mathrm{S}} \in \mathbb{R}^{T\times(N-1)}$. Therefore, the gain of power agent i at time $t=T+1,\ldots,T+H$ is measured by its marginal profit,

$$\mathcal{G}_i(x_{i,t};\mathbf{X}^{\mathrm{S}},\mathcal{M}_i) = \Big(\rho(\hat{x}_{i,t}^{\text{market}},x_{i,t}) - \rho(\hat{x}_{i,t}^{\text{local}},x_{i,t})\Big)^{+}, \quad (12)$$

where $(x)^{+}=\max(0,x)$, $\hat{x}_{i,t}^{\text{local}}=\mathcal{M}_i(\mathbf{x}_i^{\text{B(ts)}};\mathbf{x}_i^{\text{B(tr)}})$ is the forecast using only data from buyer i and $\hat{x}_{i,t}^{\text{market}}=\mathcal{M}_i(\mathbf{X}^{\text{S(ts)}};\mathbf{X}^{\text{S(tr)}})$ is the forecast obtained by combining local data and data from other agents — $\mathbf{x}_i^{\text{B(tr)}}, \mathbf{X}^{\text{S(tr)}}$ are the sets used to train the models, while $\mathbf{x}_i^{\text{B(ts)}}, \mathbf{X}^{\text{S(ts)}}$ are the sets used to forecast $\{\hat{x}_{i,t}\}_{t=T+1}^{T+H}$. By simplicity, the same model $\mathcal{M}$ and gain function $\mathcal{G}$ are used for all the buyers, but conceptually buyers may provide their own $\mathcal{M}_i$ and $\mathcal{G}_i$ to the market operator.

The last two parameters from buyers are the private valuation of gain $\mu_i \in \mathbb{R}^{+}$, i.e., a trade-off value that means how much buyer i is willing to pay for a unit increase in gain, and the public bid price $b_i \leq \mu_i, b_i \in \mathbb{R}^{+}$. Note that buyers enter the market to buy forecasts $\{\hat{x}_{i,t}\}_{t=T+1}^{T+H}$, without knowing which data were used to produce the forecasts, $H \geq 1$.

3) Market Operator: The role of the market operator includes feature allocation (Section III-C1), market price definition (Sections III-C2 and III-C4), revenue extraction from the buyers (Section III-C2) and corresponding distribution to the sellers (Section III-C3).

It is important to underline that only the market operator has access to input data (power measurements, NWP, etc.) and is responsible for fitting the quantile regression model described in Section II-B. Sellers only have access to their own time-series and buyers only have access to power forecasts produced for their power plants. Therefore, data privacy is guaranteed, assuming that the market operator is a trustworthy and neutral agent. Note that the data market framework can be applied to any forecasting methodology and the use of quantile regression is not a fundamental requirement.

B. Data Market Mechanism

At time $t = T$, RES agents provide their historical data to the market operator. Then, agent i aims to forecast the power for the next H time steps, $\{\hat{x}_{i,t}\}_{t=T+1}^{T+H}$, and the following steps occur in sequence (illustrated in Fig. 2):

S1 The marketplace sets a market price $p_i \in \mathbb{R}^{+}$ for a unit increase in gain when forecasting $\{\hat{x}_{i,t}\}_{t=T+1}^{T+H}$, following the market solution (i.e., bid and market price, forecasting accuracy) for the previous buyer $i-1$,

$$p_i = \mathcal{PF}(b_{i-1}, p_{i-1}; \Theta_{i-1}), \quad (13)$$

where $\mathcal{PF}$ is the market price update function, and $\Theta_{i-1} = (\mathcal{M}_{i-1}, \mathcal{G}_{i-1}, \mathbf{X}^{\mathrm{S}}, \mathbf{x}_{i-1}^{\mathrm{B}})$ – market operator decides p_i before buyer i arrives and according to the previous prices, otherwise truthfulness is not ensured.

S2 Buyer i bids b_i, which maximizes its value function,

$$b_i = \underset{z\in\mathbb{R}^{+}}{\operatorname{argmin}} \underbrace{\mu_i \textstyle\sum_t \mathcal{G}_i(x_{i,t};\Theta_i) - \mathcal{RF}(p_i, z; \Theta_i)}_{\mathcal{U}_i(z,\{x_{i,t}\}_{t=T+1}^{T+H}) \text{ = value function}}, \quad (14)$$

and is related to the difference between the value derived from the gain in forecasting accuracy and the data market price, $t = T+1,\ldots,T+H$. $\mathcal{RF}$ is the revenue function.

S3 The marketplace allocates available features according to the market price and bid price,

$$\tilde{\mathbf{X}} = \mathcal{AF}(p_i, b_i; \mathbf{X}^{\mathrm{S}}), \quad (15)$$

with $\mathcal{AF}$ representing the allocation function.

S4 The marketplace extracts revenue r_i from buyer i,

$$r_i = \mathcal{RF}(p_i, b_i; \Theta_i). \quad (16)$$

S5 Market divides r_i among the $N-1$ sellers using

$$\psi_i(m) = \mathcal{PD}(\mathbf{x}_i^{\mathrm{B}}, \tilde{\mathbf{X}}, K; \mathcal{M}_i, \mathcal{G}_i), \tag{17}$$

where $\mathcal{PD}$ is the payment division function, $m \in \mathcal{A}\backslash\{i\}$.

S6 Buyer i receives $\{\hat{x}_{i,t}\}_{t=T+1}^{T+H}$ and leaves the market.

S7 If a new time step occurred, sellers update their data and send it to the market operator.

C. Market Configuration

Certain properties must be met in order to produce a fair auction mechanism when defining $\mathcal{PF}$, $\mathcal{AF}$, $\mathcal{RF}$ and $\mathcal{PD}$, from (13) to (17). First, the auction mechanism needs to encourage buyers to declare their true valuation for an increase in forecasting skill. This is achieved through the *allocation* and *revenue* functions. From the other side, the market operator needs to incentivize sellers to participate in the market, meaning that the *revenue division* function should ensure three properties:

i) money paid by the buyer is totally divided by the sellers;
ii) sellers with similar information receive the same amount of money;
iii) irrelevant information receives zero payment.

1) Allocation Function: The allocation function $\mathcal{AF}(p_i, b_i; \mathbf{X}^{\mathrm{S}})$ defines the information that marketplace should use when forecasting the time-series of buyer i. The proposed mechanism assumes that all available features are used to train and evaluate the forecasting model. However, in order to ensure that the allocated features are a function of the difference between the bid price and the market price, the model is fitted (and the gain is estimated) using a perturbed version of competitors' data. More specifically, the allocated features are obtained by

$$\tilde{x}_{j,t} = \begin{cases} x_{j,t} + \max(0, p_i - b_i)\mathcal{N}(0, \sigma^2), & j \neq i \\ x_{j,t}, & j = i. \end{cases} \tag{18}$$

where $\mathcal{N}(0, \sigma^2)$ is a univariate Gaussian distribution.

2) Revenue Function: The revenue function $\mathcal{RF}(p_i, b_i; \Theta_i)$ is computed by the market operator based on its model estimation for each buyer i. The market price is based on the gain to buyer i, which is unknown for the future but can be estimated through holdout cross-validation. While forecasting $\{x_{i,t}\}_{t=T+1}^{T+H}$, the marketplace splits $\mathbf{X}^{\mathrm{S}}$ into training, validation and testing data, $\mathbf{X}^{\mathrm{S(tr)}}$ is used to estimate the model, $\mathbf{X}^{\mathrm{S(val)}}$ is used to estimate the gain and $\mathbf{X}^{\mathrm{S(ts)}}$ to forecast $\{x_{i,t}\}_{t=T+1}^{T+H}$. $\mathbf{X}^{\mathrm{S(val)}}$ corresponds to the set used to forecast the last Δ values $\{x_{i,t}\}_{t=T}^{T-\Delta+1}$, and $\mathbf{X}^{\mathrm{S(tr)}}$ to the sample used to forecast the remaining $T-\Delta$ observations $\{x_{i,t}\}_{t=1}^{T-\Delta}$, as illustrated in Fig. 3, $\Delta \geq 1$. Moreover, as previously mentioned, the data market should price the forecasts according to the marginal gain accrued to its buyers. Fig. 4 illustrates the difference between paying by the gain and paying by the marginal gain as defined by the Myerson's payment function rule [21]

$$\begin{aligned}\mathcal{RF}(p_i, b_i; \Theta_i) = b_i \mathcal{G}_i(\mathbf{x}_i^{\mathrm{B(val)}}; \mathcal{AF}(p_i, b_i; \mathbf{X}^{\mathrm{S}}), \mathcal{M}_i) \\ - \int_0^{b_i} \mathcal{G}_i(\mathbf{x}_i^{\mathrm{B(val)}}; \mathcal{AF}(z, b_i; \mathbf{X}^{\mathrm{S}}), \mathcal{M}_i) dz,\end{aligned} \tag{19}$$

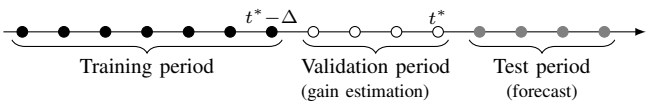
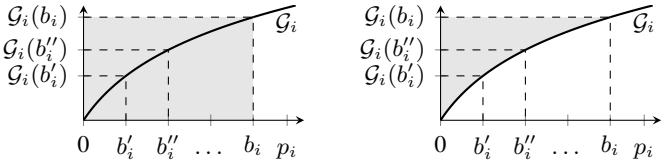


Fig. 3. Timeline for current time t^*.



(a) Based on gain ($\mathcal{G}_i$) (b) Based on *marginal* gain ($\mathcal{RF}$)

Fig. 4. Difference between paying by the gain and paying by the marginal gain (market price = shadow area, x axis = bid price, y axis = gain).

which is adopted in this paper — if bid prices b_i' and b_i'', with $b_i'' > b_i'$, might produce similar gain, $\mathcal{G}_i(b_i') \approx \mathcal{G}_i(b_i'')$, then a RES agent is incentivized to bid b_i'' anyway since it would only pay b_i' according to the marginal gain rule.

A revenue close to zero means that the buyer is purchasing low-quality forecasts, particularly when bid and market prices are high and an higher revenue from data sharing was expected.

3) Payment Division Function: The payment division function $\mathcal{PD}(\mathbf{x}_i^{\mathrm{B(val)}}, \mathcal{AF}(p_i, b_i; \mathbf{X}^{\mathrm{S}}), K; \mathcal{M}_i, \mathcal{G}_i)$ divides the value r_i paid by buyer i among the $N-1$ sellers. Ideally, the relevance of each feature would be estimated by training the statistical model $\mathcal{M}_i$ with all possible feature combinations. This method is known as Shapley Allocation [22] and ensures the three properties listed at the beginning of this section. However, when a large number of sellers is considered, this strategy may be computationally infeasible.

To overcome this challenge, the Shapley Approximation method uses a smaller number of possible feature combinations [23]. Given a random permutation σ of all features' indices $\{1, \ldots, N\}$, from an universe $\boldsymbol{\sigma}$, two models are trained using the features given by $\sigma_i < m$ and $\sigma_i \leq m$. The importance of a feature m is given by the difference in gains between these two models. The process is repeated K times and averaged out. Theoretically, the Shapley approximation $\hat{\psi}_i(m)$ achieves $\|\psi_i^{\text{shapley}}(m) - \hat{\psi}_i(m)\| < \varepsilon$, with probability $1-\zeta$ if $K > [N \log(2/\zeta)]/(2\varepsilon)^2$. Since the models are trained multiple times for different agents, the choice of the model $\mathcal{M}_i$ clearly affects the computational efficiency of the payment division function.

Furthermore, a post-processing step is applied to make the algorithm more robust to data replication. Consider a data market with three sellers, S_1, S_2 and S_3, such that S_1 and S_2 have uncorrelated and equally relevant data for buyer i, while S_3 is irrelevant, i.e $\psi_i(1)=\psi_i(2)=0.5$ and $\psi_i(3)=0$. If S_1 replicate its data once and sell again in the marketplace, the proportion of received payment will be $\psi_i(1)=2/3$, $\psi_i(2)=1/3$. Since sellers provide a unique time-series, they cannot replicate data; yet, they can collude with other agents and negotiate a portion of the extra revenue. If S_1 and S_3 collude, then $\psi_i(1)=\psi_i(2)=\psi_i(3)=1/3$.

In order to avoid data replication, the weight $\psi_i(m)$ of each

seller m is penalized if its data are similar to others in the market. This penalty is related to the cosine similarity, which measures the similarity between two vectors $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^T$ as

$$\mathcal{SM}(\mathbf{x}_1, \mathbf{x}_2) = \frac{|\langle \mathbf{x}_1, \mathbf{x}_2 \rangle|}{\|\mathbf{x}_1\| \|\mathbf{x}_2\|}, \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^T, \tag{20}$$

where $|\langle . \rangle|$ and $\|.\|$ denote the absolute value of the dot product and the Euclidean norm, respectively.

Algorithm 1 illustrates the algorithm to determine the proportion that a seller should receive from the buyer's payment. Regarding the example with three sellers, if S_1 replicates data then the Shapley allocation using Algorithm 1 (with λ=1) decreases to $1/(2+e^2) < 1/2$.

Algorithm 1 Payment division algorithm ($\mathcal{PD}$).

Input: $\mathbf{x}_i^{\mathrm{S}}$, $\tilde{\mathbf{X}} = \mathcal{AF}(p_i, b_i; \mathbf{X}^{\mathrm{S}})$, $\mathcal{M}_i$, $\mathcal{G}_i$, K
Output: $\psi_i = [\psi_i(m) \colon m \in \mathcal{A} \backslash \{i\}]$
for $m \in \mathcal{A} \backslash \{i\}$ **do**
  for $k \in \{1, \ldots, K\}$ **do**
    $\sigma_k \leftarrow \text{Uniform}(\boldsymbol{\sigma})$
  # Train models with "tr" data and forecast with "val" data
    $G = \mathcal{G}_i(\mathbf{x}_i^{\mathrm{S(val)}}; \tilde{\mathbf{X}}_{[\sigma_k < m]}, \mathcal{M}_i)$
    $G^{+m} = \mathcal{G}_i(\mathbf{x}_i^{\mathrm{S(val)}}; \tilde{\mathbf{X}}_{[\sigma_k < m] \cup m}, \mathcal{M}_i)$
    $\hat{\psi}_i^k(m) = (G^{+m} - G)^+$
  end for
  $\hat{\psi}_i(m) = \frac{1}{K} \sum_{k=1}^{K} \hat{\psi}_i^k(m)$
end for
$\psi_i'(m) = \hat{\psi}_i(m) \exp(-\lambda \sum_{j \in \mathcal{A} \backslash \{i,m\}} \mathcal{SM}(\mathbf{x}_m^{\mathrm{S}}, \mathbf{x}_j^{\mathrm{S}}))$
$\psi_i(m) = \psi_i'(m) / \sum_{m \in \mathcal{A} \backslash \{i\}} \psi_i'(m)$

4) Market Price Update Function: The function $\mathcal{PF}(b_{i-1}, p_{i-1}; \Theta_{i-1})$ computes the market price of the data for buyer i based on the gain from the other agents. We assume a set of possible market prices $\mathcal{B}_p$, which ranges from a minimum value $p_{\min}$ and a maximum value $p_{\max}$, with increment Δ_p. When the data market initializes, the market price is uniformly sampled from $\mathcal{B}_p$. Then, the market operator uses the forecasting accuracy from the first agent and estimates the revenue for each possible market price. The probabilities are updated and used to generate the market price when a new buyer arrives, iteratively, ensuring the truthfulness of the data market. Algorithm 2 proposes an online balance for the trade-off between large and small market prices. Considering a bid price b_i, if p_i is too large then the positive term in $\mathcal{RF}$ will be small (as the deterioration of $\mathbf{X}^{\mathrm{S}}$ is very high) leading to lower revenue. Similarly, if p_i is too small, the negative term in $\mathcal{RF}$ will be large, which again leads to an undesired loss in revenue.

D. Available Platforms for Implementation

This marketplace can be implemented in readily available platforms and protocols, reviewed below, which enable data transaction, verification and payment capabilities.

Ocean Protocol is an ecosystem for data trading, built on top of blockchain technology, where Oceans Tokens are used as the unit of exchange for buying or selling data services [24]. Enigma provides a protocol for secret contracts, which are similar to smart contracts but bring privacy by offloading the computation over sensitive data to an external network where it may be broken into different nodes and apply cryptographic techniques [25]. SingularityNET is a decentralized platform for trading Artificial Intelligence (AI) services, including data, through the native platform's cryptocurrency [26]. Numerai is an AI platform that aims at bringing together the best experts in data science for making forecasts for a common dataset and those who perform well are reward with some Numeraires (i.e., cryptocurrency token) and those who did not perform well will lose the Numeraires staked [27].

The majority of these platforms lack from an advanced model for data trading and, therefore, a synergy between the market mechanism described in this work and blockchain-powered platforms (e.g, tokens, protocols and smart contracts) can be established for a real-world implementation of this concept.

Algorithm 2 Market price update algorithm ($\mathcal{PF}$).

Input: b_{i-1}, p_{i-1}, $p_{\min}$, $p_{\max}$, Δ_p, Θ_i=$(\mathcal{M}_i, \mathcal{G}_i, \mathbf{X}^{\mathrm{S}}, \mathbf{x}_i^{\mathrm{B}})$
Output: p_i
$\mathcal{B}_p \leftarrow [p_{\min}, p_{\min} + \Delta_p, p_{\min} + 2\Delta_p, \ldots, p_{\max}]$
Initialize the weights for each possible market price
$w_1^j \leftarrow 1, \forall j = 1, \ldots, |\mathcal{B}_p|$
# When a buyer enters the market, the market price is determined and the weights are updated for the next buyer
for $i = 1, \ldots, |\mathcal{A}|$ **do**
  $p_i \leftarrow \mathcal{B}_p(j)$ with probability $w_i^j / \sum_{j=1}^{|\mathcal{B}_p|} w_i^j$
  for $j = 1, \ldots, |\mathcal{B}_p|$ **do**
    $g_i^j \leftarrow \mathcal{RF}(\mathcal{B}_p(j), b_i; \boldsymbol{\Theta}_i)$ # revenue for the j-th price
    $w_{i+1}^j \leftarrow w_i^j (1 + \delta g_i^j)$ # update weights
  end for
end for

IV. Case Studies

In this section, three different case studies are constructed to evaluate the proposed no-regret auction mechanism: (i) synthetic data with 3 agents aiming to verify, with a simple setup, how the data market operates; (ii) synthetic data with 50 agents, aiming to evaluate the effect of different covariance matrices in the data market (iii) wind power data publicly available from the Nord Pool electricity market.

A. Synthetic Data: Simple Setup with 3 Agents

1) Data Description and Experiments: Three agents are assumed. Let $x_{i,t}$ denote the observations from agent i at time t, and $\mathbf{x}_t$=$[x_{1,t},\ x_{2,t},\ x_{3,t}]$, where i=$1, \ldots, 3$ and t=$1, \ldots, T$. The synthetic data are generated from the VAR model,

$$\mathbf{x}_t = \begin{pmatrix} 0.5 & 0.7 & -0.1 \\ 0 & 0.7 & 0.1 \\ 0 & 0 & 0.8 \end{pmatrix} \mathbf{x}_{t-1} + \boldsymbol{\varepsilon}_t, \tag{21}$$

where $\boldsymbol{\varepsilon}_t = \begin{bmatrix} \varepsilon_{1,t} & \varepsilon_{2,t} & \varepsilon_{3,t} \end{bmatrix}$ are the error terms, $\varepsilon_{i,t} \sim \mathcal{N}(0, 1)$.

As experiments, hour-ahead forecasts are validated using an out-of-sample fold with 150 consecutive time steps. The market operator uses a sliding window with the 8760 most

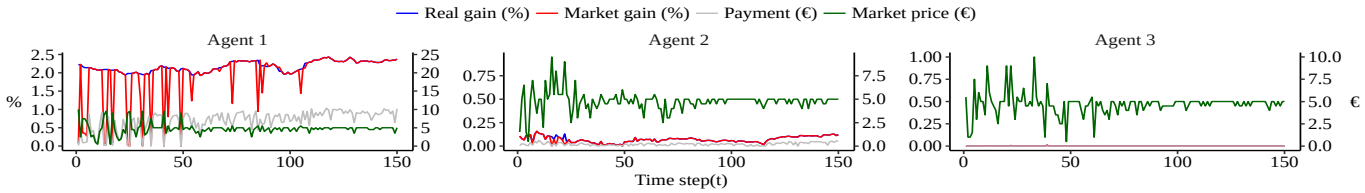


Fig. 5. Market dynamics for experiment E1 (bid price is constant and equal to 5€).

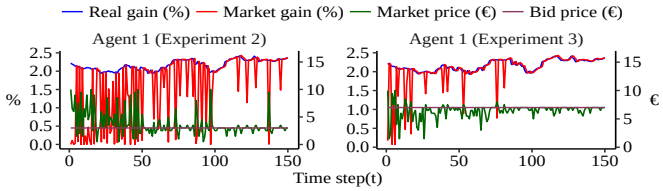


Fig. 6. Market dynamics for agent 1 in experiments E2 and E3.

recent observations divided in 8592 for model fitting and 168 to estimate the improvement in gain.

For the data market simulation, a linear regression is used as the model $\mathcal{M}_i$, $\forall i \in \{1,2,3\}$, with covariates provided by the 1h-lagged time-series. The gain function $\mathcal{G}_i$ is the improvement over the model estimated by using only its own (lagged) time-series, in terms of normalized root mean squared error (NRMSE) measured for each agent i as

$$\text{NRMSE} = \frac{\sqrt{\frac{\sum_{t=1}^{T}(\hat{x}_{i,t}-x_{i,t})^2}{T}}}{\max(\{x_{i,t}\}_{t=1}^{T}) - \min(\{x_{i,t}\}_{t=1}^{T})} \times 100. \quad (22)$$

The market operator sets a market price between 0.50€ and 10€, with 0.50€ increment, for each 1% improvement in NRMSE when forecasting one time-step ahead. The auction mechanism is simulated through the following experiments, which assume that the buyers have the following bid prices (both market and bid prices are expressed in € per 1% improvement in NRMSE):

- **E1** A fixed bid price of 5€; i.e., each agent values a marginal improvement of 1% in NRMSE as 5€.
- **E2** Agents bid fixed values of 3€, 5€ and 7€, respectively.
- **E3** Agents bid fixed values of 7€, 5€ and 3€, respectively.
- **E4** Agents bid price according to the NRMSE of their local model. Agents with a poor local model are more prone to improve 1% in NRMSE. The functional relation between the bid price and local model NRMSE is expressed as

$$b(\text{NRMSE}) = \frac{10}{1+\exp(-0.3 \times \text{NRMSE} + 5)}. \quad (23)$$

The NRMSE for the local model is estimated using the Δ most recent observations.

2) Results and Discussion: Fig. 5 depicts market dynamics when buyers always bid price 5€. At the end of 100 iterations, the market price tends to the bid's price values. As expected, when the market price is below or equal to the bid price, the gain corresponds to the gain using the real model. On the other hand, when the market price is higher than the bid price, the gain is reduced as a consequence of the noise addition into the covariates from the other agents. Furthermore, in all experiments, agent 1 has the highest benefit when using data from the market, which was expected by (21).

Additionally, since the gain for agents 2 and 3 is small, their payment is also small even when the market is not adding noise to the covariates. The payment from agent 1 is divided by agents 2 and 3 through a mean percentage of 97.6% and 2.4%, respectively, which is coherent with the fair distribution. When some gain is estimated for agent 2, agent 3 receives 100% of the value paid. Even though agent 3 does not benefit in terms of forecast accuracy improvement, it receives money from agents 1 and 2 who are not aware that agent 3 is selling data in the market.

Fig. 6 depicts data market dynamics for agent 1, at experiments E2 and E3. Since the gain to agents 2 and 3 is small, the market price is influenced by the bid price of agent 1, and the former conclusions stand. Furthermore, when the agent with the highest gain bids closer to the initial market price, the market price converges faster.

The market price and revenue dynamics for E4 (not depicted in Fig. 6) are similar to the ones from E2, where agent 1 bids at a price higher than agents 2 and 3. Since the NRMSE for the local forecasting model is stationary for all agents (with values around 9.8%, 7.7% and 5.8%, respectively), the agents bid prices around 1.2€, 0.65€ and 0.37€ per 1% improvement in NRMSE, respectively. The market price converges to 1€ per 1% improvement in NRMSE.

B. Synthetic Data: 50 Agents

1) Data Description and Experiments: Let $\mathbf{x}_t=[x_{1,t},\ldots,x_{50,t}]$. The synthetic data for the 50 agents are generated from the VAR model

$$\mathbf{x}_t = \mathbf{B}\mathbf{x}_{t-1}+\boldsymbol{\varepsilon}_t, \quad (24)$$

where $\mathbf{B}$ is the coefficient matrix, $\mathbf{B} \in \mathbb{R}^{50\times 50}$, and $\boldsymbol{\varepsilon}_t=[\varepsilon_{1,t} \ \ldots \ \varepsilon_{50,t}]$ is the error vector, $\varepsilon_{i,t} \sim \mathcal{N}(0,1), \forall i$. Two datasets ($\mathbf{D_1}$ and $\mathbf{D_2}$) are generated to evaluate the effect of different covariance matrices in the proposed approach.

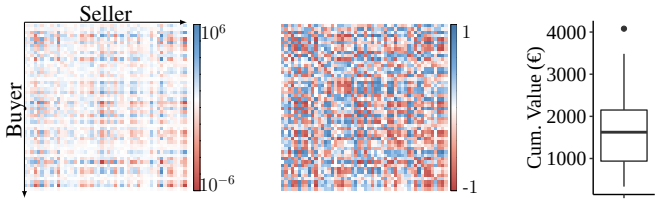
$\mathbf{D_1}$ assumes a sparse $\mathbf{B}$ matrix where: Agents $1, 2, 12, 16$, 21 and 43 should benefit with forecasts from the data market; agents $2, 3, 11, 12, 36$ and 44 should receive payment from the data market. $\mathbf{D_2}$ assumes a $\mathbf{B}$ matrix such that a large number of time-series is highly-correlated.

As in Subsection IV-A1, hour-ahead forecasts are validated using an out-of-sample fold with 150 consecutive time steps. The market operator uses a sliding window with the 8760 most recent observations divided in 8592 for model fitting and 168 to estimate the improvement in gain. $\mathcal{M}_i$ is a linear

TABLE I
CUMULATIVE GAINS WITH $\mathbf{D_1}$ BY AGENT (€).

	1	2	3	11	12	16	21	36	43	44	Others
Payment	589.3	30.0	0.0	0.1	88.2	58.4	22.7	0.0	101.4	0.0	[0,2[
Revenue*	0.6	570.8	26.6	98.7	48.9	2.3	0.3	19.1	0.5	90.0	[0,2[
Tot. Gain**	519.2	595.1	26.6	98.7	176.3	123.4	17.6	19.1	110.8	90.0	[0,4[

* Revenue = data market revenue (i.e. value received by selling data)
** Tot. Gain = data market revenue + revenue with purchased forecasts - value paid



(a) Covariance matrix. (b) Pearson correlation. (c) Total gains.

Fig. 7. Covariance and correlation for data $\mathbf{D_2}$ and gain after 150 time steps.

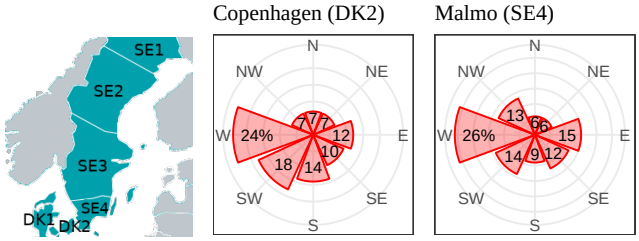


Fig. 8. Nord Pool regions in Denmark (DK) and Sweden (SE), as well as the wind roses for the wind direction observed in Copenhagen and Malmo.

regression with covariates given by the 1h-lagged time-series, and $\mathcal{G}_i$ is the improvement over the model estimated by using only its own (lagged) time-series, in terms of NRMSE. The market operator sets a market price between 0.50€ and 10€, with 0.50€ increment, and each agent values a marginal improvement of 1% in NRMSE as 5€.

2) Results and Discussion: Table I summarizes the results for $\mathbf{D1}$, at the end of 150 time steps. Sellers with data that improve the forecasts of other agents get higher revenue from the data market, when compared to the others. Conversely, agents that buy forecasts with higher accuracy pay higher values, but are compensated by the gain associated with the imbalance costs reduction. For instance, agent 1 pays 589.3€ but the extra gain from using these forecasts, instead of those obtained by its internal (or local) model, is 1107.9€.

Fig. 7 summarizes the covariance and correlation matrices for $\mathbf{D_2}$, as well as the total gain (boxplot) for the 50 agents. There is a large number of correlated time-series. But once again, agents gain money by improving their forecasting accuracy or by selling their data to others. The lowest total gain is 333.4€ and more than 30 agents receive at least 1000€.

C. Nord Pool Data

1) Data Description and Experiments: Nord Pool runs the largest market for electrical energy in Europe, operating in several northern Europe countries. For illustrative purposes, we use the historical wind power values, spot price and imbalance prices for upward and downward regulation, available in the Nord Pool website[1], from 6 regions: 4 in Sweden (SE1, SE2, SE3, SE4) and 2 in Denmark (DK1 and DK2). In this test case, each region is assumed to represent an electricity market agent. The dataset ranges between 1st January 2016 and 12th October 2017 with hourly resolution. Fig. 8 provides a geographical representation of these regions as well as the wind roses for the wind direction observed in Copenhagen and Malmo during this period[2].

[1]https://www.nordpoolgroup.com/ (accessed on June 2020)
[2]https://www.weatheronline.co.uk (accessed on June 2020)

The agents are assumed to maximize their electricity market's revenue trough forecasting the optimal quantile $\alpha_t^* = \frac{\hat{\lambda}_t^{\downarrow}}{\hat{\lambda}_t^{\downarrow} + \hat{\lambda}_t^{\uparrow}}$, as in Section II-A. Lags 1, 2 and 3 are used as covariates in the QR model provided by (11), motivated by preliminary cross-correlation analysis of the time-series. The gain is computed by the improvement in the electricity market revenue, as defined in (12), which measures how much money an agent earns on the electricity market when using the forecast provided by the data market instead of the forecasts obtained through the use of local data (and model).

As in the previous case-study, hour-ahead power forecasts are generated and validated in the same way. The parameter η, used for forecasting upward and downward regulation unit costs, is estimated (i.e., select the value with minimum mean square error) by dividing the first one-year data in 9 months for training the Holt-Winters model and the remaining 3 months for computing the corresponding mean squared error, for $\eta \in \{0.9, 0.95, 0.99, 0.999\}$.

In this test case, the market operator is assumed to set a market price between 5% and 70% of the gain, with 5% increments, i.e. for each 1€ increase in electricity market revenue, the market operator may define a market price between 0.05€ and 0.70€. On the other hand, the bid price is 50% for all buyers, i.e. the buyers are willing to pay a maximum of 0.50€ for each 1€ increase in electricity market revenue.

2) Results and Discussion: For each time step, the gain in electricity market revenue is computed as the difference between the revenue obtained when using forecasts from the data market and the revenue obtained by using a local forecasting model built without neighbor time-series. Fig. 9 depicts the cumulative revenue gain from the electricity market, i.e. the extra revenue obtained by using the forecast provided by the data market. Furthermore, the same plot shows the cumulative revenue from the data market, i.e. how much each agent receives by sharing data with the market operator, and, finally, the cumulative payment that each agent pays to the data market in order to buy forecasts. Table II supports the graphical analysis by presenting the cumulative gains and total revenue at the end of the testing period (approx. 10 months).

An agent participating in the data market may increase its revenue either by receiving more money from the electricity market (i.e., minimizing imbalance costs) or by receiving money from the data market (i.e., selling data to competitors). The fundamental goal of the data market is to have a total revenue (i.e. sum of revenues obtained in the data and electricity

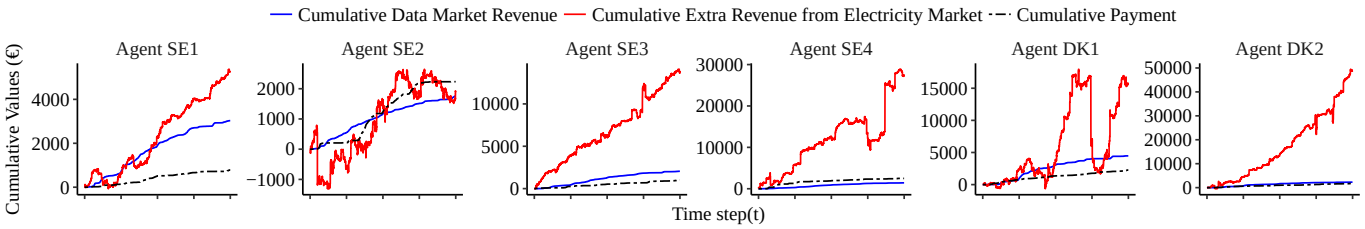


Fig. 9. Cumulative values for electricity market revenue (over a quantile regression using only local data), data market revenue and payment.

TABLE II
CUMULATIVE GAINS (€) AT THE END OF TESTING PERIOD. (1ST JANUARY 2017 TO 12TH OCTOBER 2017)

	SE1	SE2	SE3	SE4	DK1	DK2
Electricity Market	5303	1907	13668	27393	15609	48883
Paid value	783	2233	1042	2501	2271	1798
Data Market	3030	1758	2048	1429	4471	2292
Total revenue	7550	1432	14674	26321	17809	49377

TABLE III
PAYMENT DIVISION BY THE COMPETITORS (IN %).

	SE1	SE2	SE3	SE4	DK1	DK2
SE1	—	29.11	10.70	20.86	35.69	3.64
SE2	19.45	—	12.52	24.51	29.79	13.73
SE3	10.21	25.96	—	15.33	39.80	8.70
SE4	10.62	27.01	9.54	—	42.55	10.28
DK1	1.52	9.60	10.73	28.78	—	49.38
DK2	2.01	9.42	5.13	20.50	62.94	—

market minus the payment to the data market) higher than the revenue obtained in the electricity market without third-party data or data monetization.

Agent DK2 benefits the most from the data market, followed by agent SE4. These benefits are mainly due to the increase in the revenue from the electricity market, i.e. from the improvement of the forecasting models. This is explained by the fact that wind comes predominately from the West (as depicted in Fig. 8), and their forecast models are improved by the time-series from agent DK1 (located to the East).

On the other hand, the agent DK1 receives a higher reward for sharing its data with the market operator, which is also coherent with predominant wind direction. Southwest locations will be more relevant to improve forecasting models. Consequently, northwestern regions tend to benefit most from using forecasts with information from the other agents. The sudden decrease in accumulated gains (e.g., for agent DK1) occur due to extremely high values for regulation unit costs. For agent DK1, the high losses are associated with a upward regulation unit costs higher than 200€/MWh (when the values in 99% of the historical period are smaller than 30€/MWh).

Finally, Table III summarizes how the value paid by each agent is divided by the other agents (data sellers). By construction, the proportion that a data seller receives is related to the relevance (i.e., explanatory power) of its time-series when forecasting the RES generation of a buyer. Agent DK1 receives a higher reward for sharing its data, which is due to its geographical location. Following the same reasoning, it would be expected that SE1 received a smaller proportion of money from all the competitors.

In order to assess the added value of a quantile regression with varying nominal proportions over time (α_t) instead of a constant value α, the mean values for $\lambda^{\uparrow}$ and $\lambda^{\downarrow}$ are computed for the testing period and the related nominal proportion is estimated. The value for the nominal proportion is 0.60. The results show that the revenue from the electricity market for agents SE1, SE2, SE3, SE4, DK1 and DK2 increases, respectively, 176,298€, 517,218€, 437,747€, 293,813€, 887,684€ and 344,883€ when using α_t instead of α.

V. CONCLUSIONS

Data sharing between different owners has a high potential to improve RES forecasting skill in different time horizons (e.g., hours-ahead, day-ahead) and consequently the revenue from electricity market players. However, economic incentives, trough data monetization, are fundamental to implement collaborative forecasting schemes since RES agents can be competitors, and therefore unwilling to share their confidential data without benefits. This work was inspired by [16] and adapted for RES forecasting. The gain function of buyers was adapted for RES agents, which have a local model with their own variables and enter the market to improve it with more information. Furthermore, an evaluation was performed using three case studies.

Synthetic data was used in a controlled case study where it was possible to confirm: (i) the correct allocation of revenue across sellers by the market operator, and (ii) the buyers who did not benefit from the forecasts of others did not pay for such forecasts. Data from the Nord Pool market was used to evaluate the potential of a data market for RES agents, and it was concluded that: (i) all agents benefit (from the economic point of view) from the data market, (ii) agents that first observe wind-flow (or wind generation) in one location, e.g. at timestep $t-1$, provide relevant information to improve the forecasting model (e.g., for $t+1$) of neighbor agents in other locations, conditioned by wind direction, and then all agents benefit by the higher revenue accrued either from the data market or the better forecast in the electricity market. In summary, data markets can be a solution to foster data exchange between RES agents and contribute to reduce imbalance costs.

In this work, linear quantile regression and the Holt-Winters statistical models were used for the power and imbalance prices forecasts respectively. However, the choice of these models, considering aspects such as time horizon, non-linear relation between power and NWP, etc., must be carefully

considered to deliver maximum gains in the electricity and data markets. For instance, the market operator can use a statistical model tailored to each RES agent, as long as the forecasting skill is maximized since it impacts the financial incentives to share data.

In future work, the loss of RES agents when sharing their data should be considered when defining the data price. Evidently, a seller sharing data with its competitors expects a compensation for the potential impact on its business. Furthermore, some improvements are required when using a sliding-window approach. The current version of the algorithm works by adding noise to the covariates, which means that, for each new time step, the market operator needs to perform a batch train that can result in a high computational effort as more and more agents enter the market. Ideally, the noise should be introduced in the output of the model, allowing the market operator to update the model weights through online learning whenever the variables in the data market remain the same. The privacy of the data is another issue to address, since in our simulations the agents share the data with the market operator, which may represent an obstacle for some agents. Finally, another topic for future work is to develop peer-to-peer data trading schemes (i.e., without a central node as market operator) for prosumers in local energy communities, in such a way that data sellers can set their own data price.

REFERENCES

[1] J. R. Andrade and R. J. Bessa, "Improving renewable energy forecasting with a grid of numerical weather predictions," *IEEE Transactions on Sustainable Energy*, vol. 8, no. 4, pp. 1571–1580, 2017.

[2] C. Gilbert, J. Browell, and D. McMillan, "Leveraging turbine-level data for improved probabilistic wind power forecasting," *IEEE Transactions on Sustainable Energy*, In Press, 2019.

[3] J. Tastu, P. Pinson, P.-J. Trombe, and H. Madsen, "Probabilistic forecasts of wind power generation accounting for geographically dispersed information," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 480–489, 2014.

[4] R. Bessa, A. Trindade, and V. Miranda, "Spatial-temporal solar power forecasting for smart grids," *IEEE Transactions on Industrial Informatics*, vol. 11, no. 1, pp. 232–241, 2015.

[5] J. Parra-Arnau, "Optimized, direct sale of privacy in personal data marketplaces," *Information Sciences*, vol. 424, pp. 354–384, 2018.

[6] S. Mehta, M. Dawande, G. Janakiraman, and V. Mookerjee, "How to sell a dataset? Pricing policies for data monetization," in *20th ACM Conference on Economics and Computation (EC'19)*. ACM, 2019, pp. 679–679.

[7] I. Koutsopoulos, A. Gionis, and M. Halkidi, "Auctioning data for learning," in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE, 2015, pp. 706–713.

[8] X. Cao, Y. Chen, and K. R. Liu, "Data trading with multiple owners, collectors, and users: An iterative auction mechanism," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 3, no. 2, pp. 268–281, 2017.

[9] D. Acemoglu, A. Makhdoumi, A. Malekian, and A. Ozdaglar, "Too much data: Prices and inefficiencies in data markets," National Bureau of Economic Research, Inc., Tech. Rep. NBER Working Papers 26296, 2019.

[10] S. Xuan, L. Zheng, I. Chung, W. Wang, D. Man, X. Du, W. Yang, and M. Guizani, "An incentive mechanism for data sharing based on blockchain with smart contracts," *Computers and Electrical Engineering*, vol. 83, p. 106587, 2020.

[11] A. Yassine, A. A. N. Shirehjini, and S. Shirmohammadi, "Smart meters big data: Game theoretic modelfor fair data sharing in deregulated smart grids," *IEEE Access*, vol. 3, pp. 2743–2754, Dec. 2015.

[12] O. Samuel, N. Javaid, M. Awais, Z. Ahmed, M. Imran, and M. Guizani, "Blockchain model for fair data sharing in deregulated smart grids," in *2019 IEEE Global Communications Conference (GLOBECOM)*, Waikoloa, HI, USA, USA, Dec. 2019.

[13] C. Aperjis and B. A. Huberman, "A market for unbiased private data: Paying individuals according to their privacy attitudes," *First Monday*, vol. 17, no. 5–7, May 2012. [Online]. Available: https://journals.uic.edu/ojs/index.php/fm/article/download/4013/3209

[14] C. Niu, Z. Zheng, S. Tang, X. Gao, and F. Wu, "Making big money from small sensors: Trading time-series data under pufferfish privacy," in *IEEE Conference on Computer Communications (IEEE INFOCOM 2019)*. IEEE, 2019, pp. 568–576.

[15] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, p. 12, 2019.

[16] A. Agarwal, M. Dahleh, and T. Sarkar, "A marketplace for data: An algorithmic solution," in *2019 ACM Conference on Economics and Computation*. ACM, 2019, pp. 701–726.

[17] P. Pinson, C. Chevallier, and G. N. Kariniotakis, "Trading wind generation from short-term probabilistic forecasts of wind power," *IEEE Transactions on Power Systems*, vol. 22, no. 3, pp. 1148–1156, 2007.

[18] T. Soares, P. Pinson, T. V. Jensen, and H. Morais, "Optimal offering strategies for wind power in energy and primary reserve markets," *IEEE Transactions on Sustainable Energy*, vol. 7, no. 3, pp. 1036–1045, Jul. 2016.

[19] T. Jónsson, P. Pinson, H. Nielsen, and H. Madsen, "Exponential smoothing approaches for prediction in real-time electricity markets," *Energies*, vol. 7, no. 6, pp. 3710–3732, 2014.

[20] R. Koenker and K. F. Hallock, "Quantile regression," *Journal of economic perspectives*, vol. 15, no. 4, pp. 143–156, 2001.

[21] R. B. Myerson, "Optimal auction design," *Mathematics of operations research*, vol. 6, no. 1, pp. 58–73, 1981. [Online]. Available: oceanprotocol.com/tech-whitepaper.pdf

[22] L. S. Shapley, "A value for n-person games," *Contributions to the Theory of Games*, vol. 2, no. 28, pp. 307–317, 1953.

[23] S. Maleki, L. Tran-Thanh, G. Hines, T. Rahwan, and A. Rogers, "Bounding the estimation error of sampling-based Shapley value approximation," in *Fourth Workshop on Cooperative Games in Multiagent Systems (CoopMAS-2014)*, May 2014.

[24] "Ocean protocol: A decentralized substrate for AI data & services," 2019. [Online]. Available: oceanprotocol.com/tech-whitepaper.pdf

[25] G. Zyskind, O. Nathan, and A. Pentland, *New Solutions for Cybersecurity*. MIT Press, 2018, ch. Enigma: Decentralized Computation Platform with Guaranteed Privacy, pp. 425–454.

[26] B. Goertzel, S. Giacomelli, D. Hanson, C. Pennachin, and M. Argentieri, "SingularityNET: A decentralized, open market and inter-network for AIs," *Thoughts, Theories & Studies on Artificial Intelligence (AI) Research*, Dec. 2017. [Online]. Available: airesearch.com/ai-research-papers/singularitynet-a-decentralized-open-market-and-inter-network-for-ais/

[27] R. Craib, G. Bradway, X. Dunn, and J. Krug, "Numeraire: A cryptographic token for coordinating machine intelligence and preventing overfitting," 2017. [Online]. Available: numer.ai/whitepaper.pdf

Carla Gonçalves is a Ph.D. candidate in Applied Mathematics from the Faculty of Sciences of the University of Porto (FCUP), and a researcher at INESC TEC, Portugal. She received the M.Sc. in Applied Mathematics from FCUP, in 2015. Her research focuses on probabilistic and collaborative forecasting methods, with a special emphasis on renewable energies.

Pierre Pinson (SM'13, F'20) received the M.Sc. degree in applied mathematics from the National Institute for Applied Sciences, Toulouse, France, and the Ph.D. degree in energetics from Ecole des Mines de Paris, Paris, France. He is a Professor with the Centre for Electric Power and Energy, Department of Electrical Engineering, Technical University of Denmark, Lyngby, Denmark, also heading a group focusing on energy analytics and markets. His research interests include forecasting, uncertainty estimation, optimization under uncertainty, decision sciences, and renewable energies. He is the Editor-in-Chief for the International Journal of Forecasting..

Ricardo Bessa (M'18–SM'19) received the *Licenciado* (5-years) degree in electrical and computer engineering, the M.Sc. degree in data analysis and decision support systems and the Ph.D. degree in Sustainable Energy Systems (MIT Portugal) from the University of Porto. He is coordinator of the Center for Power and Energy Systems at INESC TEC. His main research interests include renewable energy, energy analytics, smart grids and electricity markets. Serves as an Editor for the *IEEE Transactions on Sustainable Energy*.