Optimization and Engineering manuscript No.
(will be inserted by the editor)

Modelling of a fully renewable energy grid with hydrogen storage: time aggregation for a scalable Capacity Expansion Problem

Bianca Urso · **Gabor Riccardi**

Received: date / Accepted: date

Abstract In recent years, the integration of renewable energy sources into electrical grids has become a critical area of research due to the increasing need for sustainable and resilient energy systems. To adress the variability of wind and solar power output over time, electricity grids expansion plans need to account for multiple scenarios over large time horizons. This significanlty increases the size of the resulting Linear Problem (LP), making them computationally challenging for large scale grids. To tackle this, we propose an approach that aggregates time steps to reduce the problem size, followed by an iterative refinement of the aggregation. We provide sufficient conditions under which the aggregated problem is equivalent to the original, unaggregated one, and refine the time intervals that do not meet these conditions. **write better:Additionally, we introduce a validation function to assess the feasibility of the aggregated solutions. Our method is tested on a fully renewable energy grid with hydrogen storage.** We generate scenarios for wind and photovoltaic (PV) power output using scenario generation techniques that capture temporal dependencies. These dependencies are modeled with marginal distributions coupled with a Gaussian copula, ensuring that the generated scenarios reflect realistic temporal correlations observed in historical data.

Keywords First keyword · Second keyword · need 4-6

Mathematics Subject Classification (2020) 90-10 · 90B15

B. Urso
IUSS School of Advances Studies, Palazzo del Broletto, Piazza della Vittoria, 15 – 27100 Pavia PV, Italy
Tel: +39 0382 375811
Fax: +39 0382 375899
E-mail: bianca.urso@iusspavia.it

G. Riccardi
Dipartimento di Matematica ”F. Casorati” Via Adolfo Ferrata, 5 – 27100 Pavia

1 INTRODUCTION

Context; literature overview. Small introduction on LP/MIP. Definition of stochastic optimization, definition of CEP, ED problems.

The threat of climate change is pushing policy-makers to pursue greater integration of renewable energy sources into electrical grids, while at the same time ensuring reliability and resilience through digital optimization of electric power distribution and transmission in smart grids [?]. One of the main difficulties arising when designing an electric power system relying on renewables is the great variability of the generation of electricity through wind and solar, since these resources are highly dependent on weather conditions, curtailment = inefficient, need hydrogen, cite source for hydrogen (eg: EU super investing in it). The stochastic nature of the problem.... making it impossible to plan long-term by optimizing on forecasts, and requiring a statistical approach to ensure a robust model. Up to now, common approaches have adopted Stochastic Programming (SP) or Robust Optimization (RO) models, along with hybrid models involving Information Gap Decision Theory or Chance Constraint [1]. While initially favored, the SP approach comes with high computational burden, so RO models have seen more popularity in recent years, despite the drawback of being conservative methods with higher average cost of operation and planning of energy systems.

In the typical setting, the problem to solve is a Capacity Expansion Problem (CEP) regarding infrastructure investments: solar and wind farms, fuel cells, hydrolizers, grid upgrades to augment Net Transfer Capacity (NTC) and so on. Nested within the CEP is an Economic Dispatch (ED) problem concerning the operational costs of said infrastructure. The problem is well suited to be modelled through mixed integer linear programming (MILP), as is explained in detail in [?] and ((other examples?))

The CEP for investment planning requires to look at long time horizons, and on the other hand intra-day variability in generation is the main complexity driver for the ED problem, so the time horizon must be modelled by a large quantity of fairly tight time steps. Furthermore, spacial structure big, even though we ignore it here and just slam everything into country nodes. Thus the temporal and spacial characteristics of the model bring the MILP size to increase rapidly. This is especially demanding in the case of SP, since all the variables from the inner ED problem must be reproduced over all scenarios.

To reduce these costs, one possible approach is to use a Rolling Horizon (RH) [explain better]. The basic technique is described in [?], along with some results regarding quality guarantees for the optimality of the solution. In [?], a rolling horizon approach is used within a RO model to optimize the operation of a micro-grid composed of check serving an isolated area in Chile. other examples of RH used for the ED A similar idea is applied in [?], where integer variables representing capital investments are initially relaxed and then progressively fixed in successive time steps, reducing the computational costs associated to the search for integer solutions.

A big drawback of the RH approach is that Problem: it is not optimal. In fact, [?]. But better reflects actual decision making based on info that is only available

progressively, vs perfect foresigh approach.

Our work aims to build a hybrid model, exploring the use of the RH method as a tool to guide progressively tighter relaxations of the perfect foresight model, rather than as a stand-alone technique. Explain better. Further, RH is then used for the validation of the results for the CEP obtained in the perfect foresight optimization, to ensure that given a solution to the CEP, the ED problem admits feasible solutions even in a limited-foresight environment. The idea is to solve the CEP as to build a grid that can be operated as a "smart grid", through a control system that optimizes on day ahead forecasts.

The model we optimize on in introduce model: components, aim, use Gurobi...

Finally, to generate the scenarios needed to train and test the SP model, a Gaussian Copula method was used. This has been previously done here and here and it's not new right?.

This paper is organized as follows: HOW?

2 MODEL

2.1 LP FORMULATION

The scenario generation step is followed by an optimization process. The model takes in input the generation and load scenarios of a given area along with various parameters indicating costs and efficiency of the current state of technology and possible upper bounds for the decision variables, and returns the optimized capacity that is necessary to meet demand throughout a one year span, with minimal cost. The optimization problem is solved using the Gurobi solver. When optimizing over multiple scenarios jointly, the solver returns the minimal amount of infrastructure and capacities that is needed to have feasibility (that is, demand met at all time - no blackouts) over all scenarios in input, with minimal average cost over the scenarios.

Our model describes a network in Europe that is to be powered and supplied of hydrogen trough power generated by photovoltaic panels and wind turbines, converted to hydrogen through electrolysis and potentially reconverted in fuel cells. It describes the handling of hydrogen throughout a one year time span, with one hour time steps, in order to meet demand.

The network is represented by an undirected multigraph $\mathscr{G} = (\mathscr{N}, \mathscr{E})$, where $\mathscr{N}$ corresponds to the nodes in the network, and $\mathscr{E} = \mathscr{E}_H \cup \mathscr{E}_P$ represents transmission lines ($e \in \mathscr{E}_P$) and hydrogen lines ($e \in \mathscr{E}_H$). Each of these nodes can be in different countries in Europe, and the power generated by wind and solar power depends on the node location. In particular, if node $n \in \mathscr{N}$ is in France, the scenarios for power generation at n will be generated using parameters fitted to France's data.

The operation of the Electric Grid is modeled as a two-stage stochastic program. In the first stage, the capacity expansion of each generator, hydrogen storage, trans-

mission line, and hydrogen pipe is decided. In the second stage, the economic dispatch is solved for each scenario.

2.2 Decision Variables

The main variables that are of interest to the policy maker and are explicitly returned in output are the following:

ns_n : Number of solar units (integer) at node $n \in \mathcal{N}$;
nw_n : Number of wind units (integer) at node $n \in \mathcal{N}$;
nh_n : Storage capacity (kg) (continuous) at node $n \in \mathcal{N}$;

Stored hydrogen is considered to be the total of liquid and gas hydrogen to be stored. Our model does not assume a distinction between the two forms, and considers hydrogen to be immediately ready for long-term storage as soon as it is converted from electricity, as well as instantaneously convertible to electricity in fuel cells at need.

To give an indication of how much hydrogen should be kept in gas form at fuel cells, ready to be converted, and how much storage for gas hydrogen should be available at electrolyzers to accomodate abundant generation moments, we consider the following decision variables as well (N.B.: they are values *per hour*, to be compared to the time necessary for gas to liquid hydrogen transformation and vice versa):

mhte_n : Maximum hydrogen to electricity capacity (Kg) (continuous), i.e. the maximum amount of hydrogen that needs to be converted within a 1h time frame into electricity at node $n \in \mathcal{N}$;
meth_n : Maximum electricity to hydrogen capacity (MWh) (continuous) at node $n \in \mathcal{N}$;

To accommodate for possible improvements on capacity of existing power lines or hydrogen transport infrastructure, the following variables are added:

addNTC_l : additional net transfer capacity on line l;
addMH_l : additional hydrogen transfer capacity on pipe l.

And parameters for the cost of such new infrastructure have to be given, indexed by edge of the respective graph: $cNTC_l$ and cMH_l.

The second stage variables, indexed by scenario j, time step t, and node $n \in \mathcal{N}$ are:

$\text{H}_{j,t,n}$: Stored hydrogen at time t, scenario j and node n (kg) (continuous);
$\text{HtE}_{j,t,n}$: Hydrogen converted to electricity at time t,scenario j (kg) (continuous);
$\text{EtH}_{j,t,n}$: Electricity converted to hydrogen at time t, scenario j (MWh) (continuous);

Note: all variables are set to be non-negative.

2.3 Parameters

There are a series of parameters that characterize the model and can be modified by the policy maker through the GUI. The main ones are related to capital costs of the infrastructure to be built:

$$\text{cs}_n : \text{Cost of one Solar Panel (€)};$$
$$\text{cw}_n : \text{Cost of one Wind Turbine (€)};$$

If not specified through the GUI, the following values are assumed for panels and turbines: cs = 400€, cw = 3 000 000€. In this model we assume no marginal costs for PV and wind power production: the operating costs of the farms throughout their life-cycle can be factored into the capital costs, and there is no additional cost linked to the production itself. We assume the flow of electricity has no marginal cost nor power loss (the modelling of that problem is beyond the scope of this project). Conversely, we do set a cost for the use of hydrogen pipes (or other means of transfer):

$$\text{cH\_edge}_l \ : \ \text{Cost of transferring 1kg of hydrogen through edge } l.$$

Conversely, the marginal costs of conversion within electrolyzers and power cells are relevant. Thus one can set the following parameters:

chte : Conversion cost of 1 kg of hydrogen to electricity (€/kg);
ceth : Conversion cost of 1 MWh of electricity to hydrogen (€/MWh).

According to the European Hydrogen Market Landcape November 2023 Report [2], "Hydrogen production costs via electrolysis with a direct connection to a renewable energy source in Europe vary from 4.18 to 9.60 €/kg H2 of hydrogen, with the average for all countries being 6.86 €/kg H2". For electrolyzers, we consider the Levelised cost of hydrogen (LCOH) to account for both marginal costs and capital costs. Such cost is dependent on the country's specific market condition and can be calculated through the European Hydrogen Observator tool. Unless specified through the GUI, $ceth$ = 20kg/MWh*10€/kg=200€/MWh is assumed (see the discussion on conversion efficiency below), and $chte$ is assumed to be 2€/kg. Furthermore, the storage of the hydrogen itself has a cost that depends on various factors: capital cost of the technology used for storage, operating costs, length of time that the hydrogen is kept in storage. Thus the following parameters can be set:

ch : Cost of hydrogen storage capacity per unit of hydrogen (€/kg);
ch_t : Cost of storing hydrogen for 1h, per unit of hydrogen (€/(kg·h)).

Thus ch will be simply multiplied by the maximum storage needed (nh), representing capital cost of storage infrastructure, whereas ch_t represents the marginal cost of keeping the hydrogen stored. Unless specified, it is assumed to be $ch = 10$€/kg and $ch_t = 0$€/(kg·h)

Within the electrolyzers and fuel cells, the conversion itself can be more or less efficient:

fhte : efficiency of hydrogen to electricity conversion (scalar between 0 and 1);

feth : efficiency of electricity to hydrogen conversion (scalar between 0 and 1).

It is assumed that 1kg of hydrogen has an energy value of 33kWh. Thus if we consider an electrolyzer operating at maximum efficiency ($feth = 1$), one MWh of electricity yields 1000/33$\simeq$30kg of hydrogen. If not specified through the GUI, a value of $feth = 0.66$ is considered, thus 1MWh yields 20kg of hydrogen. Conversely, in a fuel cell operating at maximum efficiency ($fhte = 1$) 1kg of hydrogen yields 33kWh. If not specified through the GUI, a value of $fhte = 0.75$ is considered, yielding 24.75kWh per kg of hydrogen. Actual efficiencies vary a lot depending on the technology used. Furthermore, chemical and physical constraints make it so that efficiencies higher than 0.80-0.85 are currently considered unachievable [**?**].

Finally, the GUI gives the option to place upper bounds to the variables, based on either technological and physical constraints (dimension of the facilities) or because of political choices (local population unfavourable to wind turbines):

Mns : Maximum number of solar panels that can be installed (integer);

Mnw : Maximum number of wind turbines that can be installed (integer);

Mnh : Maximum hydrogen storage capacity (kg);

Mhte : Upper bound for $mhte$ (kg);

Meth : Upper bound for $meth$ (MWh).

If no bound is given through the GUI, these values will be set to $Mns = 10^6, Mnw = 500, Mnh = 10^9, Mhte = 10^6, Meth = 10^5$). Note: computation time increases significanly when increasing the upper bound for Mnw.

A scenario consists in a different realizations of the following variables, given as input in the model and indexed by scenario j, time step t, and node $n \in \mathcal{N}$:

$\text{ES}_{j,t,n}$: Power output of a single solar panel (MWh)

$\text{EW}_{j,t,n}$: Power output of a single wind turbine (MWh)

$\text{EL}_{j,t,n}$: Electricity load (MWh)

$\text{HL}_{j,t,n}$: Hydrogen load (kg)

$\text{P\_edge}_{j,t,l}$: Power passing through line l during time step t in scenario j (MWh) (continuous);

$\text{H\_edge}_{j,t,l}$: Hydrogen flowing through pipe l during time step t in scenario j (kg) (continuous).

Lastly the following parameters are indexed by edge of the respective graph:

NTC_l : Net Transfer Capacity, that is, maximum amount of electricity that can pass on line l of the electric grid in the span of 1h;

MH_l : Maximum amount of hydrogen that can flow on edge l in 1h.

2.3.1 Objective Function

The cost function is given by the sum of all capital costs of installing infrastructure, all marginal costs of the hour-by-hour hydrogen to electricity and electricity to hydrogen conversions, and minimal costs associated to the variables *mhte* and *meth* so that they are minimized through the model. Let $Nnodes$, $NEedges$ and $NHedges$ be the number of nodes, edges on the electric grid graph and edges of the hydrogen transfer graph respectively. The objective function is modified as follows:

$$\begin{aligned}
\min \quad & \sum_{k=1}^{\text{Nnodes}} \text{cs}_k \cdot \text{ns}_k + \text{cw}_k \cdot \text{nw}_k + \text{ch}_k \cdot \text{nh}_k + \\
+ \quad & \sum_{l=1}^{\text{NEedges}} \text{cNTC}_l \cdot \text{addNTC}_l + \sum_{l=1}^{\text{NHedges}} \text{cMH}_l \cdot \text{addMH}_l + \\
+ \quad & \frac{1}{d} \sum_{j=1}^{d} \sum_{i=1}^{\text{inst}} \left(\sum_{k=1}^{\text{Nnodes}} (\text{ch_t}_k \cdot \text{H}_{j,t,k} + \text{chte}_k \cdot \text{HtE}_{j,t,k} + \text{ceth}_k \cdot \text{EtH}_{j,t,k}) + \right. \\
& \left. + \sum_{l=1}^{\text{NHedhes}} (\text{cH_edge}_l \cdot \text{H_edge}_{j,t,l}) \right) + \\
+ \quad & \sum_{k=1}^{\text{Nnodes}} 0.01 \cdot (mhte_k + meth_k)
\end{aligned}$$

The $1/d$ factor in front of the marginal costs allows to average over the scenarios, whereas the capital costs are the same for all scenarios. Thus, ignoring the costs of *mhte* and *meth*, the objective function value gives an estimate of the actual costs (in €) for the set up and maintenance of the system throughout the length of the scenario (one year).

2.4 Constraints

The following constraints are to ensure that for all time steps t and all scenarios j, the electricity load and the hydrogen load are met. The measure units are MWh and kg respectively, conversion factors are considered for HtE and EtH respecively. Let $Out(n)$ and $In(n)$ indicate the outgoing and incoming edges from node n on the respective graph. Then for each node n, the we have the following flow balance constraints:

$$\begin{aligned}
\text{Electricity Balance:} \quad & \text{ns}_n \cdot \text{ES}_{j,t,n} + \text{nw}_k \cdot \text{EW}_{j,t,n} + 0.033 \cdot fhte_k \cdot \text{HtE}_{j,t,n} - \text{EL}_{j,t,n} - \text{EtH}_{j,t,n} + \\
& \sum_{l \in Out(n)} \text{P_edge}_{j,t,l} + \sum_{l \in In(n)} \text{P_edge}_{j,t,l} \geq 0; \\
\text{Hydrogen Storage:} \quad & \text{H}_{j,t+1,n} = \text{H}_{j,t,n} + 30 \cdot feth_k \cdot \text{EtH}_{j,t,n} - \text{HtE}_{j,t,n} - \text{HL}_{j,t,n} - \\
& - \sum_{l \in Out(n)} \text{H_edge}_{j,t,l} + \sum_{l \in In(n)} \text{H_edge}_{j,t,l}
\end{aligned}$$

We ask that the consumed electricity be less or equal than the produced electricity at all times. On the grid itself, the two sides should be equal, but we observe that $ns \cdot ES_{j,t} + nw \cdot EW_{j,t}$ indicate the maximum power that can be generated with set weather conditions, whereas actual production will be regulated to meet demand through curtailment.

The stored hydrogen at time $t+1$ is the result of what was stored at time t adjusted by what was converted and what was sent to the industrial load. For $t = 24 * 365$ we set the same constraint on hydrogen by considering $t+1$ to be index 1: this way we avoid placing a "start time" at an arbitrary place within the year (time is rendered modulo the year) and we avoid the model asking for conveniently high initial storage values of hydrogen appearing out of thin air.

The total storage and conversion capacities are calculated by minimizing the maximum over time and scenarios of the variables $H_{j,t}, EtH_{j,t}$ and $HtE_{j,t}$, for all scenarios j, time steps t and nodes n:

$$\begin{aligned} \text{Storage Capacity Limit:} &\quad \mathrm{H}_{j,t,n} \leq \mathrm{nh}_n; \\ \text{EtH Conversion Limit:} &\quad \mathrm{EtH}_{j,t,n} \leq \mathrm{meth}_n; \\ \text{HtE Conversion Limit:} &\quad \mathrm{HtE}_{j,t,n} \leq \mathrm{mhte}_n. \end{aligned}$$

Finally, edge capacities on the respective graphs are considered for all scenarios j, time steps t and nodes n:

$$\begin{aligned} \text{Net Transfer Capacity:} &\quad |\mathrm{P\_edge}_{j,t,l}| \leq \mathrm{NTC}_l + \mathrm{addNTC}_l; \\ \text{Hydrogen Transfer Capacity:} &\quad |\mathrm{H\_edge}_{j,t,l}| \leq \mathrm{MH}_l + \mathrm{addMH}_l. \end{aligned}$$

3 Optimization and Time Resolution

The time horizon generated by the scenarios has a time resolution where each time step has a length of one hour. Each value represents the total power (hydrogen) production or demand in the corresponding hour at the node. The smaller the length of each time step, the more accurate the results. However, the number of variables and constraints grows linearly with the number of time steps, making the model intractable (especially in the context of an application) with just a few scenarios.

Moreover, considering every hour in each day of the year is partly redundant, as each day will be similar to neighboring days. Yet, simply considering a sample of days for each season might undermine long-term storage capacity representation.

Given an initial time horizon $\mathscr{T} = \{1, \ldots, T\}$, we can consider partitions of $\mathscr{T}$ as a family of disjoint subsets whose union is $\mathscr{T}$. We only consider those partitions where every subset is an interval of $\mathscr{T}$. We refer to these as time partitions. Given a time partition P, we can consider the corresponding model obtained by considering each interval in P as a single time step. For every I in P, we define:

$$ES_{j,I,n} \coloneqq \sum_{i \in I} ES_{j,i,n}, \quad EW_{j,I,n} \coloneqq \sum_{i \in I} EW_{j,i,n}$$

and similarly for $HL_{j,I,n}$ and $HR_{j,I,n}$. We denote the model obtained by the time partition P as CEP_P.

It is evident that the optimal value of CEP_P is a lower bound for the original problem $CEP_{\mathscr{T}}$, as given a feasible solution $(ns, nw, nh, mhte, meth, H, HtE, EtH, Pedge, Hedge)$ of the latter, we can obtain a solution of the former by taking $(ns, nw, nh, mhte, meth)$ the same as in $CEP_{\mathscr{T}}$ and:

$$Pedge_{j,I,e} = \sum_{i \in I} Pedge_{j,i,e}, \quad Hedge_{j,I,e} = \sum_{i \in I} Hedge_{j,i,e}$$

and similarly for EtH and HtE, and $H_{j,I,n} = H_{j,i_0,n}$ where $I = [i_0, \dots, i_{|I|}] \in P$. In particular, there is a cost-preserving linear map from the feasible space of $CEP_{\mathscr{T}}$ to the feasible space of CEP_P, making the latter a relaxation of the former.

This is generally true when considering any time partition P' finer than P, where for every $t' \in P'$, there exists $t \in T$ such that $t' \subset t$. In particular, we have the following observation:

Observation 3.1 *Let $V_{\mathscr{P}} \subset \mathbb{R}^{N_{\mathscr{P}}}$ and $V_{P'} \subset \mathbb{R}^{N_{P'}}$ be the space of feasible solutions of $CEP_{\mathscr{P}}$ and $CEP_{P'}$, respectively. There exists a linear map $L_{P'P} : \mathbb{R}^{N_{P'}} \to \mathbb{R}^{N_P}$ such that $L(V_{P'}) \subset V_P$ and $c_P(L(x)) = c_{P'}(x)$, where c_P is the cost function of CEP_P and $c_{P'}$ is the cost function of $CEP_{P'}$.*

Thus, by iteratively solving finer time partitions, we converge to the optimal solution of $\mathscr{P}$.

3.1 Variables and Constraints aggregation

TODOS:

- When is the cost of the aggregated solutions equal to the cost of the original problem respect to the corresponding unaggregated solutioon? done
- When is the corresponding unaggregated solution optimal for the original problem? done
- How to estend obs so that it holds for ED.
- rimuovere che combinazioni devono essere convesse, fa casino e non cambia niente done
- How to get a scenario for which the aggregated ED solution is feasible
- How to easily compute when an aggregated interval is ”far” from such fesible scenario
- Unaggregate on those intervals first. done
- Add examples
- If the non aggregated variables are mixed integers, everything works the same.
- For ρ_r constraints which behave well under generalisez convex combinations of variables can be excluded. done

GR: Secondo me ci sta scriverlo in maniera un filo generale perchè: Così questo metodo non è ristretto a solo (ed esattamente) a questo modello, magari diventa chiaro che aggiungendo generatori tradizionali tutto funziona comunque e top. Oppure per problemi totalmente diversi ma con una struttura simile. Poi così alcune dimostrazioni si "semplificano", ad esempio qui il problema della conservazione dei costi nel caso le variabili sono negative non sorgeva perchè ci si è ristretti a una formulazione con variabili positive (alla quale ci si può poi ricondurre come abbiamo visto).

Varying time aggregation can be viewed as performing row and column aggregation on the original linear programming (LP) model. Consider the following general linear problem:

$$\min_{x \in \mathbb{R}^n} c^T x \tag{1}$$

$$\text{s.t.} \quad Ax = b \tag{2}$$

$$x \geq 0 \tag{3}$$

Here, A is an $m \times n$ matrix. Now, let $\sigma = \{S_1, S_2, \ldots, S_{\tilde{n}}\}$ be a partition of $[n]$ (the columns) and $\delta = \{R_1, R_2, \ldots, R_{\tilde{m}}\}$ a partition of $[m]$ (the rows), corresponding to a partition of the rows and columns of A.

We obtain the corresponding aggregated problem by replacing each set S in σ with a single row, and each set R in δ with a single column. One way to aggregate a set of rows (or columns) is by taking a convex combination of the rows (or columns), known as *weighted aggregation*.

The corresponding aggregated LP problem becomes:

$$\min_{\tilde{x} \in \mathbb{R}^{\tilde{n}}} \tilde{c}^T \tilde{x} \tag{4}$$

$$\text{s.t.} \quad \tilde{A}\tilde{x} = \tilde{b} \tag{5}$$

$$\tilde{x} \geq 0 \tag{6}$$

where $\tilde{A}$ is a $\tilde{m} \times \tilde{n}$ matrix.

In the problem under consideration, we have various types of constraints: Electricity Balance, Hydrogen Balance, Hydrogen Storage, and bounds on the variables. Given a time partition P, we define σ and δ such that each set $S \in \sigma$ corresponds to all constraints of the same type, scenario, and time index t that falls within the same time interval in T as P. Similarly, the variables (such as Power generation, Hydrogen generation, etc.) are partitioned in δ based on the same criteria.

Rows and columns are combined via equal-weight aggregation. This aggregation maintains the structure of the original problem, meaning that had we formulated the model directly with the aggregated time steps, we would have arrived at the same model. We refer to this as a *structure-preserving aggregation*, which is defined as follows:

Definition 3.2 Given an LP problem (1), we say that a weighted aggregation with respect to partitions σ, δ is *structure-preserving* if for each $R \in \sigma$ and each $r \in R$, there exists $f^r : [\tilde{n}] \to [n]$ such that:

1. $f^r|_{\text{supp}(\tilde{A}_{R,S})} : \text{supp}(\tilde{A}_{R,S}) \to \text{supp}(A_{r,f(S)})$ is a bijection such that

$$\tilde{A}_{R,C} = A_{r,f^r(C)} \text{ for all } C \in \text{supp}(\tilde{A}_{R,S})_{>1}$$

2. $f^r|_{\delta_{=1}} = Id_{\delta_{=1}}$
3. $f^{r'}(C') = f^r(C)$ then $C = C'$
4. $\{f^r(C)\}_{r \in R \in \sigma_{>1}, C \in \delta_{>1}} = \cup_{C \in \delta_{>1}} C$

GR: Per alcune oss le seguendi condizioni si possono togliere: $f^{r'}(C') = f^r(C) \implies C = C'$ e $\{f^r(C)\}_{r \in R \in \sigma_{>1}} = \cup_{C \in \delta_{>1}} C$. Usiamo entrambe le ipotesi aggiuntive più avanti, la prima ci dice che se mandiamo du variabili C, C' del problema agregato nella stessa variabile, allora le due variabili iniziali sono la stessa. La seconda che a ogni variabili non agregata corrisponde una variabile aggregata.

Where if F is a family of sets we denote as $F_{=k}, F_{>k}$ respectively the sets of F with size equal to k and greater than k. In particular $\text{supp}(\tilde{A}_R)_{>1} \subset \delta$ is the set of indices corresponding to sets $R \in \delta$ of size greater than 1. And $\text{supp}(A_r)_{>1}$ is the set of indices such that $r \in R \in \delta$ with R of size greater than 1. This implies that the coefficients of the aggregated variables in the aggregated problem match those in the original problem for the corresponding unaggregated variables, f^r can be seen as a function mapping the aggregated variables to variables of the same "type" in the unaggregated constraint. While obtaining a feasible solution to (1) from (4) is not always guaranteed, it is possible under certain assumptions. Here, if B is a matrix having I, J as set of indexes for the rows and columns respctively, then, for all $I' \subset I$ and $J' \subset J$, we denote with $B_{I',J'}$ the submatrix of B havins rows in I' and columns in J'.

Observation 3.3 *If (σ, δ) is a structure-preserving aggregation, let $R \in \sigma$ and $r \in R$. Let $\tilde{x}$ be a solution to the aggregated problem* (4)*. If $\tilde{b}_r - \tilde{A}_{R,\delta_{=1}} \tilde{x}_{\delta_{=1}} \neq 0$, define $\rho_r := \frac{b_r - A_{r,\delta_{=1}} \tilde{x}_{\delta_{=1}}}{\tilde{b}_r - \tilde{A}_{R,\delta_{=1}} \tilde{x}_{\delta_{=1}}}$. If $A_{r,\delta_{=1}} = 0$ and $b_r = 0$, then ρ_r can be chosen arbitrarily.*

If $\rho_r \geq 0$ and $x \in \mathbb{R}^n$ satisfies $x_{\delta_{=1}} = \tilde{x}_{\delta_{=1}}$ and $x_{f^r(C)} = \rho_r \tilde{x}_C$ for all $C \in \text{supp}(\tilde{A}_R)$, then x satisfies the constraint $A_r x = b_r$ of the original problem.

Proof From the hypothesis and the definition of structure-preserving aggregation, we have:

$$\begin{aligned} A_r x &= \sum_{i \in \text{supp}(A_r)} A_{r,i} x_i \\ &= \sum_{S \in \text{supp}(\tilde{A}_R)_{>1}} A_{r,f(S)} x_{f(S)} + \sum_{j \in \delta_{=1}} A_{r,j} x_j \\ &= \sum_{S \in \text{supp}(\tilde{A}_R)_{>1}} \tilde{A}_{R,S} \rho_r \tilde{x}_S + \sum_{j \in \delta_{=1}} A_{r,j} x_j \end{aligned}$$

By the definition of ρ_r:

$$\begin{aligned}\rho_r \sum_{S\in \text{supp}(\tilde{A}_R)_{>1}} \tilde{A}_{R,S}\tilde{x}_S &= \rho_r(\tilde{A}_R\tilde{x} - \tilde{A}_{R,\delta_{=1}}\tilde{x}_{\delta_{=1}}) \\ &= \rho_r(\tilde{b}_R - \tilde{A}_{R,\delta_{=1}}\tilde{x}_{\delta_{=1}}) \\ &= b_r - A_{r,\delta_{=1}}\tilde{x}_{\delta_{=1}}\end{aligned}$$

Thus, we obtain:

$$A_r x = b_r$$

□

Observation 3.4 *Let $\tilde{\rho}_r \in \mathbb{R}$ for all $r \in R \in \sigma$ be the weights of the row aggregation. Let ρ_r for $r \in R \in \sigma_{>1}$ be defined as in 3.3, if $\tilde{\rho}_r \geq 0$. If ρ_r is well defined for all $r \in R$, then we have, for all $R \in \sigma_{>1}$:*

$$\tilde{\rho}_R^T \rho_R = 1 \tag{7}$$

Proof

$$\tilde{\rho}_R^T \rho_R = \sum_{r\in R} \tilde{\rho}_r \rho_r = \frac{\sum_{r\in R} \tilde{\rho}_r (b_r - A_{r,\delta_{=1}}\tilde{x}_{\delta_{=1}})}{\tilde{b}_r - \tilde{A}_{R,\delta_{=1}}\tilde{x}_{\delta_{=1}}} = 1$$

A structure-preserving aggregation does not inherently guarantee feasibility for all constraints in the original problem. However, Observation 3.3 illustrates how to partially reconstruct a solution x for a specific constraint r by appropriately scaling the aggregated variables within the support of A_r.

Definition 3.5 For a structure-preserving, row and column aggregation (σ, δ). A constraint $r \in \sigma_{=1}$ is ρ-agnostic if for all $C \in \delta_{>1}$ such that $C \cap \text{supp}(A_{r,c}) \neq \emptyset$,

$$A_{r,f^{r'}(C)} = \tilde{\rho}_{r'} \tilde{A}_{R^{(C)},C} \text{ for all } r' \in R^{(C)}$$

with $R^{(C)} \in \sigma_{>1}$ such that $C \in \text{supp}(A_{R^{(C)}})$.

Observation 3.6 *Let x be as in Obs 3.4. If $r \in \sigma_{=1}$ is a row-agnostic constraint, then $A_r x = b_r$.*

Proof Since $\tilde{\rho}_R^T \rho_R = 1$, we have:

$$A_r x = \sum_{C\in\delta_{=1}} A_{r,C}x_{r,C} + \sum_{C\in\text{supp}(A_r)_{>1}} \sum_{r'\in R^{(C)}} A_{r,f^{r'}(C)} x_{f^{r'}(C)} = \sum_{C\in\delta_{=1}} \tilde{A}_{r,C}\tilde{x}_{r,C} + \sum_{C\in\text{supp}(A_r)_{>1}} \sum_{r'\in R^{(C)}} \tilde{\rho}_{r'}\rho_{r'}\tilde{x}_{R,C} = \tilde{b}_r = b_r$$

We can now define the hypergraph associated to the aggregation (σ, δ).

Definition 3.7 The *hypergraph associated to the aggregation* (σ, δ) is the hypergraph $\mathcal{N}, \mathcal{E}$ having as nodes the aggregated variables $\mathcal{N} := \delta_{>1}$ and as edges the subsest of $\mathcal{N}$ that appear together in not row-agnostic constraints.

When two edges (constraints) in the hypergraph, r and r', share aggregated variables, the scaling factors ρ_r and $\rho_{r'}$ must be equal to maintain consistency. Then if ρ_r can be defined consistently, by applying observation 3.3, to all $r \in R \in \sigma_{>1}$ we can construct a feasible solution for the unaggregated problem (**??**). Thus we have:

Proposition 3.8 *If (σ,δ) is a structure-preserving aggregation. Let $\tilde{x}$ be a solution to the aggregated problem* (4)*. If for all $r \in R \in \sigma_{>1}$ such that $\tilde{b}_r - \tilde{A}_{R,\delta_{=1}}\tilde{x}_{\delta_{=1}} = 0$ we have $A_{r,\delta_{=1}} = 0$ and $b_r = 0$. Define $\rho_r := \frac{b_r - A_{r,\delta_{=1}}\tilde{x}_{\delta_{=1}}}{\tilde{b}_R - \tilde{A}_{R,\delta_{=1}}\tilde{x}_{\delta_{=1}}}$ for all $r \in R \in \sigma_{>1}$ such that $\tilde{b}_r - \tilde{A}_{R,\delta_{=1}}\tilde{x}_{\delta_{=1}} \neq 0$. If $\rho_r \geq 0$ and is constant over the connected components of the hypergraph associated to (σ,δ). Then $x_{\delta_{=1}} := \tilde{x}_{\delta_{=1}}$ and $x_{f^r(C)} := \rho_r \tilde{x}_C$ for all $C \in \operatorname{supp}(\tilde{A}_R)$ and $C \in \delta_{>1}$ is well defined and is feasible solution for the unaggregated problem* (1)*.*

Observation 3.9 *Let x, $\tilde{x}$ be as defined as in proposition 3.8. If $\tilde{\rho}_r \tilde{c}_C = c_{f(r,C)}$ for some $r \in R \in \sigma_{>1}$. Then the cost of $\tilde{x}$ for the aggregated problem is equal to the cost of x in the unaggregated problem.*

Proof Let $\tilde{x}$ be a solution to the aggregated problem (4). Using observation 3.4, for all $C \in \delta_{>1}$ the cost corresponding to the variable $\tilde{x}_C$ is

$$\tilde{c}_C \tilde{x}_C = \tilde{c}_C \sum_{r \in R} \tilde{\rho}_r \rho_r \tilde{x}_C = \sum_{r \in R} \tilde{c}_C \tilde{\rho}_r \rho_r \tilde{x}_C = \sum_{r \in R} c_{f(r,C)} x_{f(r,C)}$$

Which correspond to the cost of the variables $\{x_{f(r,C)}\}_{r \in R}$. Thus

$$\tilde{c}\tilde{x} = \sum_{C \in \delta_{=1}} \tilde{c}_C \tilde{x}_C + \sum_{C \in \delta_{>1}} \tilde{c}_C \tilde{x}_C = \sum_{C \in \delta_{=1}} c_C x_C + \sum_{C \in \delta_{>1}} \sum_{r \in R} c_{f(r,C)} x_{f(r,C)} = \sum_{C \in \delta_{=1}} c_C x_C + \sum_{j \in \cup_{C \in \delta_{>1}} C} c_i x_i = cx$$

While row aggregation of a linear problem is a relaxation of the original problem, the same does not apply to column aggregation. However, the column aggregation used for the Capacity Expansion Problem in this work is still a relaxation. In general a column aggregation of a linear problem is a relaxation of the original problem whenever it is a *constant-coefficients column aggregation*, that is:

Definition 3.10 A column aggregation of a linear problem with weights $\tilde{\rho}_c$, $c \in C \in \delta$ is a *constant-coefficients column aggregation* if for all $c \in C \in \delta$ for all rows $r \in [m]$, we have $\tilde{\rho}_c A_{r,c} = \frac{1}{|C|}$ or $\tilde{\rho}_c A_{r,c} = 0$.

That is if for every set $C \in \sigma$ of variables that are aggregated together, each variable in C has the same coefficients in every row of the aggregated problem defined by σ and the coefficient is equal to the aggregation weight of the corresponding variable, except for those rows where all the coefficients of the variables are zero. We then substitute the columns corresponding to C with a vector containing one in every row in which the coefficients in C are non zero, otherwise we substitute with zero.

Proposition 3.11 *If (σ,δ) is a structure-preserving, constant-coefficients column aggregation, if the hypothesis of proposition 3.8 holds then the aggregated problem* (4) *is exact and x is an optimal solution.*

Proof Since for observation 3.9, the cost of the aggregated problem is equal to the cost of x in the unaggregated problem, we only need to show that the aggregated problem is a relaxation of the unaggregated problem. Let x be a solution to the unaggregated problem (1). For all $C \in \delta_{=1}$ let $\tilde{x}_C := x_C$. Since $\{f^r(C)\}_{r \in \cup_{R \in \sigma_{>1}}, C \in \delta} =$

$\cup_{C\in\delta_{>1}} C$, for all $c \in C \in \delta_{>1}$, exists $r \in R \in \sigma_{>1}$ and $C \in \delta_{>1}$ such that $f^r(C) = c$. Then let $\tilde{x}_C := \sum_{c\in C} A_{r,c}x_c$. Since if $f^r(C) = f^{r'}(C')$ implies that $C = C'$, x is well defined. Lastly for all $R \in \sigma$ we have:

$$\tilde{A}_R\tilde{x} = \sum_{r\in R}\left(\sum_{C\in\delta_{=1}} \tilde{\rho}_r A_{r,c}\tilde{x}_C + \sum_{C\in \mathrm{supp}(\tilde{A}_r)_{>1}} \tilde{x}_C\right) = \sum_{r\in R}\left(\sum_{C\in\delta_{=1}} \tilde{\rho}_r A_{r,c}x_C + \sum_{C\in \mathrm{supp}(\tilde{A}_r)_{>1}}\sum_{c\in C} A_{r,c}x_c\right) = \sum_{r\in R}\tilde{\rho}_r b_r = \tilde{b}_R$$

□

4 APPLICATION TO CAPACITY EXPANSION PROBLEM

We now apply the results of the previous section to the Capacity Expansion Problem. TODO:

- What are the connected components of the hypergraph as defined in the previous sectors? each connected component looks like the hypergraph of the ED at a fixed timestep. Why can we not consider the edges corresponding to hydrogen storage? because however ρs are chosen, they hold, so they don't force ρ to be equal over the nodes it connects.
- What does it mean? Given an aggregated problem, we are interested in which time intervals are problematic, since the ones which have well define ρ could be disgragated and obtain the same solution, we disgragate the ones in which the ρs are far from being well defines, that is we have two constraints in the same connected components with really different ρ_r.
- What is ρ_r, since the aggregated variables are only second stage variables. Let's consider ρ_r with $r \in R$, and let t_r and $n \in \mathcal{N}$ be respectively the time step and the node corresponding to the constraint r.Let the net energy production at time t in n: that is demand in n minus any renewable power produced in n. Then ρ_r corresponds to fraction between the net energy production at time t and the net energy production during the interval T with $t \in T$.
- How to disaggregate: Thus we calculate $\rho_(n,t)$ for all nodes in the network and al timestep, and disagregate thos with highest variance for fixed t. (since, fixed t the should all be equal for the solution be feasible for the problem obtained by disagregating the time interval T).

- How much to disaggregate: One could either subdivide each time interval into finer time intervals, or to the single timesteps. For the former, one could keep the intervals in which the variance of ρ_r is small, together, and divide in single timesteps the rest.
- Initial aggregation: to calculate ρ_r we need to have solved already an aggregated problem, so it's not too insightful. But we can start by grouping together intervals by putting "peaks" on the extremes, that is peak of demand and production, and then subdividing equally the so obtained intervals.
- Write well how to calculate ρ_r

4.1 SCENARIO GENERATION

To estimate the optimal capacities for the CEP through a stochastic approach, realistic and diverse weather scenarios are needed, so to capture the variability and uncertainty of power generation through renewable sources over extended periods. In order to generate such scenarios, samples are extracted from a joint probability density function (PDF) fit on historical data. In the following subsection, we use Y_t to denote the stochastic process of generated power observations for either solar or wind in a single country. In our project, we used an hourly time step (T=$\{1...8760\}$) and fit the wind and solar distributions separately for each country considered. To model the marginal probability distributions corresponding to the power output of wind turbines for each hour of the year, a Weibull distribution was used, justified by its proven effectiveness in capturing the variability and skewness of wind power distributions [?]. For solar power, Beta distributions were employed, as in [?].

To fit our model, we used a dataset containing 30 years of data for various European countries, which was collected by [3]. On the other hand, electricity load is taken from the ENTSO-E Statistical Reports. In this simple model, while fitting on historical data we did not account for possible changes in future climate, since the focus lies mostly in the computational aspect.
To account for interdependence between temporally near time steps, we coupled these distributions using a Gaussian Copula approach, which captures the dependencies between hourly power outputs effectively. This approach accurately mimics common weather phenomena: The Gaussian Copula represents well the coupled behavior in renewable stochastic systems [?].

A possible improvement of the generation process could be to fit wind and PV data jointly in the copula step, potentially also including load scenarios with the generation scenarios through the same approach. This would consider dependence between Energy Demand and weather conditions, but it would necessitate of the historical dataset provided for the corresponding grid, and would also further increase computational costs.

4.1.1 Parametric Estimation of Wind Power distribution

The parameters defining the Weibull Distribution are estimated using the Maximum Likelyhood Estimation (MLE). The Weibull density function is given by:

$$f(x;\theta,\gamma) = \left(\frac{\gamma}{\theta}\right) x^{\gamma-1} \exp\left(-\left(\frac{x}{\theta}\right)^{\gamma}\right)$$

where $\theta, \gamma > 0$ are the scale and shape parameters, respectively. Given observations $X_1, \ldots, X_n$, the log-likelihood function is:

$$\log L(\theta,\gamma) = \sum_{i=1}^{n} \log f(X_i \mid \theta,\gamma)$$

The optimum solution is found by searching for the parameters for which the gradient is zero :

$$\frac{\partial \log L}{\partial \theta} = -\frac{n\gamma}{\theta} + \frac{\gamma}{\theta^2}\sum_{i=1}^{n} x_i^{\gamma} = 0 \tag{8}$$

Eliminating θ, we get:

$$\left[\frac{\sum_{i=1}^{n} x_i^{\gamma}\log x_i}{\sum_{i=1}^{n} x_i^{\gamma}} - \frac{1}{\gamma}\right] = \frac{1}{n}\sum_{i=1}^{n}\log x_i \tag{9}$$

This can be solved to get the MLE estimate $\hat{\gamma}$. This can be accomplished with the aid of standard iterative procedures such as the Newton-Raphson method or other numerical procedures. This is done with the aid of the package *scipy*. Once $\hat{\gamma}$ is found, $\hat{\theta}$ can be determined in terms of $\hat{\gamma}$ as:

$$\hat{\theta} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^{\hat{\gamma}}\right)^{\frac{1}{\hat{\gamma}}} \tag{10}$$

4.1.2 Parametric Estimation of Solar Power distribution

To estimate the α and β parameters defining the Beta distribution Y, we use the Method of Moments. The mean of the random variable Y can be expressed as $\mathbb{E}[Y] = \frac{\alpha}{\alpha+\beta}$ and the variance as $Var[Y] = \frac{\alpha+\beta}{(\alpha+\beta)(\alpha+\beta+1)}$. In particular by explicating β in the first equation and substituting it in the second equation we obtain that:

$$\begin{cases} \alpha = \mathbb{E}[X]\left(\frac{\mathbb{E}[X](1-\mathbb{E}[X])}{\mathrm{Var}[X]} - 1\right) \\ \beta = (1-\mathbb{E}[X])\left(\frac{\mathbb{E}[X](1-\mathbb{E}[X])}{\mathrm{Var}[X]} - 1\right) \end{cases} \tag{11}$$

By substituting the mean and the variance with their empirical approximation we obtain the Method of Moments estimator for α and β.

4.1.3 Parametric Copula Estimation

The cumulative density function of both the Weibull and Beta distributions are continuous and invertible. Therefore, the random variables $U_t := F_{Y_t}(Y_t)$ have a uniform distribution over $[0,1]$. The copula of the random variables $\{Y_t\}_{t\in T}$ is defined as the function $C : [0,1]^T \to [0,1]$ such that

$$C(F_{Y_1}(y_1),\ldots,F_{Y_T}(y_{|T|})) = P(Y_1 \le y_1,\ldots,Y_{|T|} \le y_{|T|}). \tag{12}$$

This function always exists because of Sklar's Theorem cite Sklar?. For a given correlation matrix Σ, the Gaussian Copula with parameter matrix Σ is defined as

$$C_{\Sigma}^{\mathrm{Gauss}}(u_1,\ldots,u_T) := \Phi_{\Sigma}(\Phi^{-1}(u_1),\ldots,\Phi^{-1}(u_T)),$$

where Φ, Φ_{Σ} are the cumulative distribution functions of Gaussian variables having distribution $\mathcal{N}(0,1)$ and $\mathcal{N}(\mathbf{0},\Sigma)$ respectively. In particular if $C_{\Sigma}^{\mathrm{Gauss}}$ is the copula associated with the random variables $\{Y_t\}_{t\in T}$ then we have that the random variables

$Z_t = \Phi^{-1}(F_{Y_t}(Y_t)) = \Phi^{-1}(U_t)$ have joint distribution equal to $\mathcal{N}(0,\Sigma)$. This follows from:

$$\begin{aligned} P(Z_1 \leq z_1, \ldots, Z_T \leq z_t) &= P(\Phi^{-1}(U_1) \leq z_1, \ldots, \Phi^{-1}(U_T) \leq z_T) = \\ &= P(U_1 \leq \Phi(z_1), \ldots, U_T \leq \Phi(z_T)) = \\ &= C_\Sigma^{\text{Gauss}}(\Phi(z_1), \ldots, \Phi(z_t)) = \\ &= \Phi_\Sigma(z_1, \ldots, z_T) \end{aligned}$$

In particular, given the realization $\{y_{t,j}\}_{t\in t, j\in J}$ of the variables $\{Y_t\}_{t\in T}$, an unbiased estimation of the parameter matrix Σ is the empirical covariance matrix $\hat{\Sigma}$ of the samples $\{\Phi^{-1}(\hat{F}_{Y_t}(y_{t,j}))\}_{t\in T, j\in J}$, where $\hat{F}_{Y_t}$ is the estimated marginal distribution of the variable Y_t as seen in subsection 4.1.1 and subsection 4.1.2.

Finally, we can generate samples from a Multivariate Gaussian random variable $(Z_t, t \in T)$ having distribution $\mathcal{N}(0,\hat{\Sigma})$. Then the power output scenarios are obtained from these samples by following the previous steps backwards, that is, for each sample, computing $\hat{F}_t^{-1}(\Phi(Z_t))$ for all $t \in T$.

5 COMPUTATIONAL RESULTS

5.1 SINGLE NODE NETWORK

First, an electrical grid with a single node is considered (corresponding ideally to an area with uniform weather conditions, highly connected at low cost). A first section will consider realistic parameter combinations and describe the results given by the solver, conducting a parameter sensitivity analysis. A second section will describe a validation function that checks the results of the capacity expansion problem for feasibility on new scenarios. Concurrently, a cost function is designed to give more realistic cost estimates compared to the optimal value given by the solver.

5.2 MULTIPLE NODE NETWORK

Results are computed for a multiple node network, with additional edge variables and parameters. When considering a network with more than a single node, computational costs increase rapidly. Thus in the first section, a small analysis is carried out to determine acceptable time steps on which time dependent data can be aggregated (the gathered data is usually on hourly steps) while maintaining the quality of the solution. Some examples are then considered, and the network dynamics that arise with the introduction of edge variables are described. A mixed approach is then used to design a validation function that can deal with the complexity arising from the introduction of the network structure in the model.

6 CONCLUSIONS

References

1. Michal Jasinski, Arsalan Najafi, et al, Operation and Planning of Energy Hubs Under Uncertainty - A Review of Mathematical Optimization Approaches, 2022
2. Author, European Hydrogen Market Landscape - November 2023 Report, page numbers. Publisher, place (year)
3. Stefan Pfenninger and Iain Staffell. “Long-term patterns of European PV output us- ing 30 years of validated hourly reanalysis and satellite data”. In: Energy 114 (2016), pp. 1251–1265. issn: 0360-5442. doi: https://doi.org/10.1016/j.energy.2016.08. 060.