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 · More

Mathematics Subject Classification (2020) MSC code1 · MSC code2 · more

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.

1. 100 parole contesto sociale (è importante perché...): cita: qualche report/programma dell'UE.
2. literature review: $\sim$ 10 articoli
   - robust vs stochastic optimization, cite something
   - time aggregation: cite {**???**} for theory?
   - validation function is dealt with RH: cite{Glomb} for quality guarantees, cite {Palma-Behnke} for microgrid ED planning, maybe someone else
   - scenario generation: cite sources down
3. cosa facciamo noi di speciale? 150 parole: We want to solve CEP to build grid that can be operated as smart with rolling horizon. Often used robust approach because computationally easier, we want to use stochastic.
4. how is this paper organized?

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, 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 [**?**]. 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 ... with a nested Economic Dispatch (ED) problem concerning the operational costs of said infrastructure. Can be solved through mixed integer linear programming (MILP) ((examples: isolated case study MILP [**?**])) The computational costs are high because – requires to look at long time horizons modeled in short time frames. – especially demanding for SP since multiple 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 –Chile. is it relevant?– 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. 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.

OLD: 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 of the grids [**?**]. The model presented in this report explores the possibility of an electrical grid powered entirely through wind and photovoltaic (PV) systems, and supported by hydrogen storage. It is of interest to estimate the power generation capacity for both wind and solar, as well as the hydrogen storage and conversion capacities, that would be necessary in order to power a reliable grid supplying residential electricity load and industrial base load of both electricity and hydrogen for a given area, while minimizing the cost of implementing such infrastructure. The main difficulties arising when designing such a system lie in the great variability of the generation of electricity through wind and solar, since these resources are highly dependent on weather conditions, making it impossible to plan long-term by optimizing on forecasts, and requiring a statistical approach to ensure a robust model. A first step for estimating the optimal capacities is thus to have realistic scenarios on which to evaluate our model. To generate the scenarios for wind and solar power generation we sampled from a joint probability density function (PDF) that was fit on historical data. The construction of the PDF for wind and solar is explained in detail in subsections **??** and **??** respectively. In this simple model, while fitting on historical data we did not account for possible changes in future climate.
On the other hand, electricity load is taken from the
hrefhttps://www.entsoe.eu/data/power-stats/ENTSO-E Statistical Reports. We normalised the 2023 data by country to indicate the trend of load throughout the year, dividing by mean hourly load: this is then to be multiplied by the mean load of the area the policy maker is interested in serving with the modeled grid. The scenario generation step is followed by an optimization process. The model, as described in section **??**, takes in input the generation and load scenarios of the selected countries 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. When considering a network with more than a single node, computational costs increase rapidly. Thus 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. Results are then evaluated for a single node network and for more complex networks based in the European Union, for which the necessary data was publicly available. Appropriate validation functions are discussed to check feasibility over new scenarios, along with cost functions that give more realistic cost estimates compared to the optimal value given by the solver.

2 MODEL

2.1 SCENARIO GENERATION

To estimate the optimal capacities through a stochastic approach, realistic weather scenarios (and thus generation scenarios) are needed. To generate the scenarios for wind and solar power generation, samples are extracted from a joint probability density function (PDF) that was fit on historical data. The construction of the joint PDF for wind and solar is done through a Gaussian Copula approach. A first section will explain the theory behind this approach, and a second section will show the application in the case of weather conditions in European countries. Weibull distributions are used as marginals to model wind, and beta distributions for solar power.

Wind and solar power are inherently intermittent and uncertain, posing a challenge to their successful integration into the energy system. Hydrogen storage and other forms of energy storage offer potential solutions to mitigate these issues. However, the amount of long-term storage required in a fully renewable grid is heavily influenced by the stochastic behavior of wind and solar power. Moreover, historical data typically covers only a limited number of climate years, which restricts the ability to test the grid over long time horizons encompassing various climatic conditions. To address this limitation, we adopted a scenario generation (SG) method based on historical data of each of the European countries considered, allowing us to create realistic and diverse scenarios that better capture the variability and uncertainty of renewable energy sources over extended periods.

To model the probability distribution corresponding to the power output of wind turbines for each hour of the year, we utilized a Weibull distribution, justified by its proven effectiveness in capturing the variability and skewness of wind power distributions [[**?**]]. For solar power, a Beta distribution was employed in [[**?**]]. controllare che siamo coerenti con tutte le cose che citiamo 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.

2.1.1 Stochastic Processes description

The stochastic processes of power observations will be denoted as Y_t. Where $t \in T$, is the set indexing all the random variables which want to be considered jointly. We assume that the random variable Y_t has either a Weibull distribution, in the case of Wind Power, or a Beta distribution in the case of Solar Power.

2.1.2 Parametric Estimation of Wind Power distribution

The parameters defining the Weibull Distribution are estimated using the Maximum Likelyhood Estimation. 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 :

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

Eliminating θ, we get:

$$\left[\frac{\sum_{i=1}^{n} x_i^{\gamma} \log x_i}{\sum_{i=1}^{n} x_i^{\gamma}} - \frac{1}{\gamma}\right] = \frac{1}{n} \sum_{i=1}^{n} \log x_i \tag{2}$$

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:

$$\hat{\theta} = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^{\hat{\gamma}}\right)^{\frac{1}{\hat{\gamma}}} \tag{3}$$

2.1.3 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:

$$\begin{cases} \alpha = \mathbb{E}[X]\left(\frac{\mathbb{E}[X](1-\mathbb{E}[X])}{\mathrm{Var}[X]} - 1\right) \\ \beta = (1-\mathbb{E}[X])\left(\frac{\mathbb{E}[X](1-\mathbb{E}[X])}{\mathrm{Var}[X]} - 1\right) \end{cases} \tag{4}$$

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

2.1.4 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

$$C(F_{Y_1}(y_1),\dots,F_{Y_T}(y_{|T|})) = P(Y_1 \le y_1,\dots,Y_{|T|} \le y_{|T|}). \tag{5}$$

This function always exists because of Sklar's Theorem. The Gaussian Copula represents well the coupled behavior in renewable stochastic systems [[**?**]] and is the one used in this project. For a given correlation matrix Σ, the Gaussian Copula with parameter matrix Σ is defined as $C_\Sigma^{\text{Gauss}}(u_1,\dots,u_T) := \Phi_\Sigma(\Phi^{-1}(u_1),\dots,\Phi^{-1}(u_T))$. Where Φ, Φ_Σ are the cdf Gaussian variables having distribution $\mathcal{N}(0,1)$ and $\mathcal{N}(0,\Sigma)$ respectively. In particular if C_Σ^{Gauss} is the copula associated 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:

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

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 **??** and subsection **??**.

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 $t \in T$.

2.1.5 Generation

In our project, we used an hourly time step (T=$\{1...8760\}$) and fit the wind and solar distributions ~~separately (thus limiting the computational costs)~~jointly (TO DO). To fit our model, we used a dataset containing 30 years of data for various European countries, which was collected by [[**?**]]. For more complex versions of the model, where we considered multiple nodes at the same time, we fit (TO DO) a joint distribution considering the involved countries, capturing typical correlations of the Northern Atlantic Oscillation... TO DO (image)

We observed that the bottleneck of the Scenario Generation algorithm is the Singular Value Decomposition (SVD) of the covariance matrix $\hat{\Sigma}$. Consequently, the

computation time changes marginally with the number of generated scenarios d. Thus, in the GUI, we stored the pre-computed SVD matrices for the European countries we worked with (individually), giving the option to rapidly generate a desired amount of scenarios for those countries. We also give the option to fit new distributions for other areas by inputting one's own historical data, but we would advise doing so on larger time steps to limit the computational time.

A possible extension could be to also include load scenarios jointly 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.

2.2 LP OPTIMIZATION

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.

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 $\mathscr{T} = \{1, \dots, T\}$, 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:

$$\min_{x \in \mathbb{R}^n} c^T x \tag{6}$$

$$\text{s.t.} \quad Ax = b \tag{7}$$

$$x \geq 0 \tag{8}$$

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:

$$\min_{\tilde{x} \in \mathbb{R}^{\tilde{n}}} \tilde{c}^T \tilde{x} \tag{9}$$

$$\text{s.t.} \quad \tilde{A}\tilde{x} = \tilde{b} \tag{10}$$

$$\tilde{x} \geq 0 \tag{11}$$

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 (**??**), 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 (**??**) from (**??**) 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 (**??**). 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* **??**, *if $\tilde{\rho}_r \geq 0$. If ρ_r is well defined for all $r \in R$, then we have, for all $R \in \sigma_{>1}$:*

$$\tilde{\rho}_R^T \rho_R = 1 \tag{12}$$

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 **??** 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* **??**. *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 **??**, 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 (**??**). 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 (**??**).*

Observation 3.9 *Let $x, \tilde{x}$ be as defined as in proposition **??**. 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 (**??**). Using observation **??**, 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 **??** holds then the aggregated problem (**??**) is exact and x is an optimal solution.*

Proof Since for observation **??**, 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 (**??**). 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:

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

□

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

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