UNIVERSITÀ
DEGLI STUDI
DI PADOVA

UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Ingegneria Industriale DII
Corso di laurea magistrale in Energy Engineering

Design and operation of energy systems under uncertainty: a comparison between deterministic and stochastic approach

Relatore
Prof. Andrea Lazzaretto

Correlatore
Prof. George Tsatsaronis

Laureando
Alessandro Pampado
2039090

Anno Accademico 2022-2023

Riassunto

L'obiettivo di questa tesi è l'analisi delle fasi di design e operation di un sistema energetico con incertezze. Nel dettaglio, i risultati devono spiegare in quale misura la modellazione dell'incertezza associata all'irradianza solare e alla temperatura ambiente possa consentire il miglioramento delle scelte di design, come ad esempio una maggiore precisione riguardo la taglia di un'unità.

L'introduzione delle incertezze risulta importante a causa di diversi fattori, quali il cambiamento climatico, condizioni di mercato inaspettate, evoluzione della richiesta energetica o pianificazioni interattive. Molti studi hanno evidenziato i vantaggi legati all'analisi delle incertezze rispetto ad un tradizionale approccio deterministico. Per le fasi di design e operation di un sistema energetico, le condizioni climatiche, il prezzo dei vettori energetici e la domanda energetica sono i principali parametri incerti da tenere in considerazione.

In questo lavoro, solo le condizioni climatiche sono considerate fonti di incertezza, in modo da vedere quanto esse possano influenzare le soluzioni di design. Per affrontare il problema, si è analizzato un sistema multi-energy residenziale: l'idea è di essere nell'anno 2010, con l'obiettivo di trovare la miglior soluzione per il "futuro", corrispondente al periodo 2010-2020, usando dati storici inerenti al periodo 2005-2009. Diversi modelli deterministici e stocastici a due stadi, con riferimento a tale sistema, sono stati sviluppati per comparare le soluzioni ottimizzate con quella di riferimento per il periodo 2010-2020.

Per prima cosa, viene discusso il peso della temperatura ambiente nel processo di clustering: questo parametro è raramente considerato nella Letteratura, ma consente di migliorare la qualità della rappresentazione del dataset iniziale. Infatti, la considerazione della sola irradianza solare presenta, in media, il 10% in meno di elementi ben posizionati rispetto al processo che utilizza sia irradianza che temperatura come attributi.

In seguito, l'attenzione è posta sui diversi metodi per la generazione di giornate rappresentative, corrispondenti al periodo di ottimizzazione, per vedere qual è il più adatto ad essere utilizzato per la fase di design di un sistema energetico. Tecniche di clustering sono comparate con profili stagionali o mensili medi. La generazione di cluster stagionali è altresì discussa. I profili medi sono dimostrati essere i peggiori, presentando errori relativi fino al 13% per la funzione obiettivo, paragonata alla soluzione di riferimento. I cluster annuali performano meglio se il numero di giorni rappresentativi è basso, uguale a 4 o 8, o alto, pari a 28.

Infine, è presentata una procedura innovativa di clustering a due stadi per la generazione di scenari stocastici per i diversi giorni rappresentativi. L'idea è quella di assegnare un set di scenari di irradianza e temperatura a ciascun giorno rappresentativo. In ogni caso, le soluzioni ottenute sono troppo conservative, il che è coerente con la teoria dello stochastic programming, ma comporta costi totali elevati.

Abstract

The aim of this thesis is to study the design and operation phases of an energy system under uncertainty. In particular, results should explain whether modelling the uncertainty associated with global solar irradiance and air temperature helps improving design choices, such as components sizes.

The importance of introducing uncertainty is related to many aspects, such as climate change, unexpected market conditions, evolution of energy demand, interactive planning. Many studies highlight the advantages of uncertainty analysis with respect to traditional deterministic approaches. For the design and operation of an energy system, climate conditions, price of energy carriers and energy demand are the main uncertain parameters.

In the following work, only climate conditions are considered as a source of uncertainty, to see how much they can affect design solutions. To address the problem, a residential multi-energy system is considered: the idea is to be in the year 2010, trying to find the best solution for the "future", the period 2010-2020, using historical data from the period 2005-2009. Deterministic and two-stage stochastic models are developed, with respect to such system, to compare the optimised solutions with the reference one for the period 2010-2020.

First, the relevance of the air temperature in the clustering process is discussed: this parameter is rarely considered in the Literature, but it allows to improve the quality of the dataset representation. In fact, the clustering process with just global solar irradiance presents, as average, 10% fewer well-positioned elements than the process using irradiance and air temperature.

Then, attention is put on different methods for generating representative days as optimisation period, to see which is the most suitable to use for the design phase of an energy system. Clustering techniques are compared with average seasonal and monthly profiles. Generation of seasonal clusters is also discussed. Average profiles are proved the worst ones, presenting relative errors up to 13% for the objective function, with respect to the reference solution. Annual clusters are better than seasonal ones when the number of representative days is low, equal to 4 or 8, or high, equal to 28.

Finally, an innovative two-step clustering procedure to generate scenarios for representative days is presented. The idea is to assign a set of scenarios to each representative day. However, obtained solutions are too conservative, which is consistent with stochastic programming theory, but entails higher total costs.

Index

Nomenclature

Acronyms

APC	Announced pledges case
ASV	Average silhouette value
BPS	Building Performance Simulation
CHP	Cogeneration system
COP	Coefficient of performance
CPU	Climate policy uncertainty
DES	Distributed energy system
DSM	Demand side management
DBA	Dynamic time warping barycentre averaging
DTW	Dynamic time warping
ED	Euclidean distance
EES	Electric energy storage
EPF	Electricity price forecast
MAE	Mean absolute error
MES	Multi-energy system
NOCT	Nominal operating cell temperature
NPV	Net present value
NZE	Net-zero emission by 2050 scenario
O&M	Operation and maintenance
PDF	probability distribution function
PV	Photovoltaic system
RCP	Representative concentration pathway
RMSD	Root mean square difference
SBD	Shape based distance
SOC	State of charge
SOP	State of power
SSD	Within cluster sum of square difference
STEPS	Stated Policies Scenario
TES	Thermal energy storage
TSAM	Time series aggregation module

Greek letters

α	Amortization factor
δ_{inv}	Binary variable associated with the inclusion of a given energy conversion unit
$\delta_{unit,t}$	Binary variable associated with ON/OFF status of a given energy conversion unit at time t
$\vartheta_{unit,t}$	Additional variable to avoid bilinear constraints for a given component at time t
η_{BOS}	Efficiency for the balance of system
η_{unit}	Efficiency of a given energy conversion unit
μ_k	Centroid of a cluster k

Symbols

A_{PV}	Area of the photovoltaic system $[m^2]$
$a_{sim}(\mathrm{i})$	Average similarity of element i with respect to the all other elements of the cluster it belongs [-]
b_0	Correction factor for the efficiency of a PV system [1/°C]
$b_{dis}(i)$	Average similarity of element i with respect to the elements of the closest cluster it does not belong to [-]
C_{unit}	Size of a component of the energy system [kW]

c_{unit}	Specific cost for a given unit [€/kW]
c_f	Correction factor [-]
$E_{EES,t}$	Energy level for the electric energy storage [kWh]
$\mathbb{E}$	Expected value
$F_{unit,t}$	Fuel required by a given component at time t [sm^3/h]
f	Objective function, minimization of the total cost [k€]
G_{tot}	Global solar irradiance for the tilted plane [W/m^2]
M	Number of scenarios [-]
M_{BIG}	Parameter required for the big M method [-]
N	Number of representative days [-]
$P_{unit,t}$	Power produced by a given energy conversion unit, or required by the load, at time t [kW]
P_t^+, P_t^-	Power charged to or discharged from the electric energy storage, at time t [kW]
p_s	Probability of a given scenario [-]
$Q_{unit,t}$	Heat flow produced by a given energy conversion unit, or required by the load at time t [kW]
Q_t^+, Q_t^-	Heat flow charged to or discharged from the thermal energy storage, at time t [kW]
$Q_{TES,t}$	Level of energy for the thermal energy storage at time t [kWh]
$\mathbb{R}^n{}_+$	N-dimensional space with all non-negative real numbers
r	Interest rate
$sil(i)$	Silhouette value for a given element [-]
T_{air}	Air temperature [°C]
T_{cell}	Cell temperature [°C]
v	Wind speed [m/s]
v_{cut-in}	Cut-in velocity [m/s]
$v_{cut-out}$	Cut-out velocity [m/s]
v_{nom}	Nominal velocity [m/s]
w_d	Weight of a given representative day [-]
$\mathbb{Z}^p{}_+$	P-dimensional space of non-negative integer numbers

Subscripts

a	Attribute for the clustering process
d	Day of the year
des	Design
fix	Fixed
HN	Here and now
h	Hour of a day
i	Row of a given matrix
j	Column of a given matrix
k	Cluster
op	Operation
s	Scenario
t	Total time
var	Variable
y	Year

1. Introduction

1.1 Impact of uncertain parameters in the optimization of design and operation of energy systems

The goal of this work is to study the design and operation phases of an energy system under uncertainty. In particular, results should explain whether the modelling of uncertainty associated with some parameters helps improving design choices, such as components sizes.

Uncertainty can be defined as "any deviation from the unachievable ideal of completely deterministic knowledge of the relevant system" [1]. Clearly, several aspects of an energy system are affected by uncertainty but, first, motivations for studying it have to be discussed. Indeed, the measure of changes in performances associated with variations of parameters is more relevant than uncertainty itself. According to [2], four reasons are crucial for this study:

1. In free market conditions, unexpected energy price changes, energy carrier substitution or energy demand change cause increases in uncertainty. As reported from [3], "the difference between the summer peak demand in 1983 and that projected for 1983 a decade earlier was equivalent to the output of 300 large nuclear plants, representing an investment of about $750 billion at 1984 prices".
2. Different natural, technological, social and institutional processes and their interactions have to be considered in planning for the long term, which makes planning and modelling very complex and their output uncertain.
3. Climate change, increasing environmental restrictions and the resulting high share of intermittent energy resources imply the use of uncertainty analysis.
4. Interactive planning involves different planning participants with different worldview, interests and uncertainty perception.

Additionally, several works focus on the necessity of modelling uncertainties. For instance, [4] underlines how adopting a deterministic approach in an uncertain environment forces the designer to take precise assumptions to get accurate data, which is often hard to do.

Consequently, it is necessary to take into account what are the relevant uncertain parameters, how they vary (Section 1.2) and the interactions among them (Section 1.2.4). First, according to [4,5], a distinction must be done about nature of uncertainties:

1. Uncertainty due to variability (aleatory uncertainty). Even if increased knowledge may help on defining a scheme, it will not reduce its complexity.
2. Uncertainty due to a lack of knowledge (epistemic uncertainty), due to, for instance, lack of observations/measurements, conflicting evidence, reducible ignorance, etc.

Another distinction is done in [5], between subjective uncertain parameters, related to the experience of the designer, and objective ones, as natural gas price or load demand.

In an energy system, main uncertainty parameters can be included in:

- Climate conditions.
- Price of energy carriers.
- Energy demand curves.
- Investment costs.
- Technical performances, such as efficiencies or COPs.

In order to get an optimal solution, all these aspects should be analysed. However, it is hard to consider all of them simultaneously, mainly because of computational burdens. Therefore, the Literature helps understanding which of them can be assumed as deterministic, in order to simplify the problem without losing too much quality of the solution.

Even if most of the papers choose some of these parameters as uncertainty sources without explaining why other ones are not relevant, some works, such as [5,6], analyse in detail the importance of all of them. Zhang

et al. [5] study impact of twelve uncertain parameters, such as generators and grid efficiencies, COPs, prices of energy carriers, design capacities, to assess the convenience of a Distributed Energy system (DES) with respect to a traditional one (CES). Despite this, in the base case they proceed with the design starting from deterministic curves. Therefore, what they are doing is a sensitivity analysis (including the design components), examining which parameters are the main drivers for variations in output results, instead of adopting uncertainty analysis, whose aim is to see variations in output given by uncertain input parameters. Clearly, under these assumptions, an increase of grid efficiency entails a higher use of it, as well as an increase in the size of absorption chillers implies a lower total cost up to a certain value.

On the contrary, Mavromatidis et al. [6] try to understand which uncertainty parameters are the most relevant for design and operation of Distributed Energy Systems (DES). Differently from [5], they perform global sensitivity analysis, by using Morris Sobol methods, showing that variations in investment costs and technical performances are not relevant for the final result.

If technical performances are investigated, it is crucial to distinguish between uncertainty related to their present or future values. Indeed, as indicated in [5], many authors present works with different values for efficiencies or COPs. By the way, it is not worth to consider them as uncertain because, once a component is chosen, its technical performances are set too. Then, if a "better" component is found, it will be sufficient to correct these values. Moreover, future technical performances are complex to be evaluated, because future trends are difficult to be forecasted. Hence, because of these considerations, technical performances evaluation can be neglected.

A similar discussion can be done for design costs. In fact, present values are strictly related to the specific component, while future trends are unsuitable to be predicted, causing inaccurate solutions. Furthermore, if the design occurs just at time zero, there is no interest in a future evolution of components costs, unless another design phase is taken into account, as done in Dominguez et al. [7]. If there is not a change in information, there is not advantage in considering more than two stages for the optimization (Section 1.2.6).

Consequently, just weather conditions, prices of energy carriers and demand curves are analysed as sources of uncertainty.

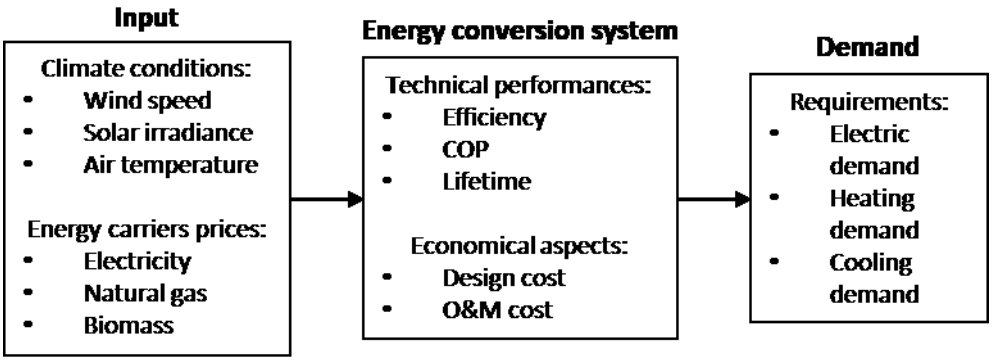

Figure 1.1.1: Uncertainty sources in an energy system.

1.2 State of the art

This section describes how uncertain parameter are modelled in the Literature. In particular, the relevant ones, according to Section 1.1, are treated: weather conditions in Section 1.2.1, prices of energy carriers in Section 1.2.2 and energy demand in Section 1.2.3. Then, interactions among uncertain parameters is discussed in Section 1.2.4. Furthermore, Section 1.2.5, about clustering techniques, and Section 1.2.6, deepening theory about stochastic programming, introduce relevant topics for the thesis.

1.2.1 Weather conditions

These are the most spread analysed uncertainty sources in the Literature. For example, [12-22] take into account them, even though they do it in different ways.

Plaga et al. [8] present an interesting and complete review about climate uncertainty in energy system models, including both deterministic and stochastic optimization problems as references. In particular, climate conditions affect many parameters.

First, wind power, sensible to wind velocity. To see the influence of wind velocity in wind turbines power production, the first step is to convert wind velocity measured at a reference height to the one obtainable at the desired weight. Then, a correction factor c_f is evaluated as a function of the velocity v.

$$c_f \begin{cases} 0 & \text{if } v<v_{cut\text{-}in} \text{ or } v>v_{cut\text{-}out} \\ \dfrac{v^3-v_{cut\text{-}in}^3}{v_{nom}^3-v_{cut\text{-}in}^3} & \text{if } v_{cut\text{-}in}<v<v_{nom} \\ 1 & \text{if } v_{nom}<v<v_{cut\text{-}out} \end{cases} \qquad (1.1)$$

The term c_f is multiplied by the nominal power to get the power production. v is the wind speed at time t, $v_{cut\text{-}in}$ is the cut-in velocity, v_{nom} the nominal velocity and $v_{cut\text{-}out}$ the cut-out one.
Anyway, this equation is not always precise, because it entails a pitch control system, which is usually adopted just for large size wind turbines because of its relevant cost. The use of curves given by manufacturers would be better. Reference [9] collects databases, studies temporal variability assessment and presents forecasting models for wind speed; [10] reviews probability distribution functions (PDF) for wind data collection; [11] compares statistical approaches for wind speed fitting. Discussion should deal with wind direction too.

Then, global solar irradiance and air temperature are crucial parameters too. Photovoltaic plants (PV) power production depends on both these parameters, through equations (1.2-1.4):

$$P_{PV} = A_{PV} * \eta_{BOS} * \eta_{PV} * G_{tot} \qquad (1.2)$$

$$\eta_{PV} = \eta_{PV}^{*}\left(1 - b_0(T_{cell} - 25°C)\right) \qquad (1.3)$$

$$T_{cell} = T_{air} + \frac{NOCT - 20}{800} * G_{tot} \qquad (1.4)$$

Equation (1.4) finds the cell temperature as a function of the air temperature, the global solar irradiance on the tilted plane and the NOCT, a parameter related to the specific panel. (1.3) finds the efficiency of the PV system as a function of the standard one, the cell temperature and a factor b_0. Finally, the power production is obtained through equation (1.2).

Additionally, other parameters are influenced by weather condition:

1. Energy demand, especially heating and cooling requirements. They can change a lot depending on the temperature.
2. Hydropower generation. Actually, instead of dealing just with technical constraints, the feasible stream flow has always to be calculated. This means that the upstream flow rate is not constant, as well as the evaporation flow. These differences are particularly relevant with a run-of-river hydropower plant, compared to a large system with a dam. Stream flow can be evaluated by using hydrological models.
3. Thermal power plants use cooling water so, if the air temperature increases, the water one will do it as well, implying a lower power plant efficiency. Moreover, water scarcity, linked to the evaporation rate, should be considered as a possible problem for the cooling system.

Therefore, climate data collection and analysis is crucial. Unfortunately, design problems deal with lots of years and, especially if stochastic programming (Section 1.2.6) is used, computational effort is enormous, due to the high number of scenarios.

Several techniques are presented in the Literature to deal with weather conditions.

- Creation of probability distribution functions. For example, in [12] solar irradiance has a beta distribution. All uncertain parameters are considered independent from each other and Montecarlo simulation is used to get final scenarios. In [13], PDFs are created for the solar irradiance, for then using discrete approximation.
- In [14], the design of the PV and off-shore wind turbine system is based on the electricity balance between their production and the electric demand. Then, production at 2050 is forecasted with machine learning techniques, obtaining PDFs, which are approximated through Markov chains Monte Carlo method.
- Han et al. [15] develop a model, using regression and Markov chains Monte Carlo method, to capture the intra and inter day variations.
- Perera et al. [16] take curves from historical data for 30 years, using Simulink. Then, they are modelled to obtain a one year probability distribution for the renewable energy generation.
- In [17], authors use clustering techniques to generate a discrete number of scenarios, starting from historical data. In particular, it is normalized and divided for winter and summer block. Then, number of representative periods is defined, as required by k-means, to generate a limited number of clusters, whose centroid is the representative day. The probability of each scenario is equal to the number of elements in each cluster divided for the total number of elements.
- Zheng et al. [18] use clustering techniques too. k-means is the adopted one to generate clusters, whose number is chosen according to the elbow method.
- In [19], authors use power curve point forecast with historical forecast errors. They create, as suggested by [20], more than 500 scenarios. To reduce computational effort, scenario reduction techniques are used.

Therefore, raw data is taken from historical databases, or evaluated through machine learning techniques. After that, a possibility is to create probability distribution functions: the main issue is the computational effort by using continuous distribution [21], forcing to simplify it through techniques as discrete approximation or Monte Carlo method. Alternatively, data series reduction is possible by generating clusters: several techniques, as k-means or k-medoids, can be used to generate proper scenarios, whose representative element is the centroid (or the medoid). Clearly, the Literature presents combinations of these two possibilities too. For example, Mavromatidis et al. [21] generate multiple weather files and, by using Monte Carlo method, 1440 scenarios, related just to weather condition, are considered. Thus, a k-medoids clustering process follows to reduce them. Finally, scenarios about extreme weather conditions are added.

Another crucial feature is whether to introduce climate changes. Liu et al. [14] do it by adopting different machine learning regression learners based on the historical weather data of the typical meteorological year. For instance, they use linear regression or regression trees. The basic data is taken from a local observatory.

Wild et al. [22] show on their paper forecasts of climate conditions, such as global solar radiation, beam solar radiation, temperature and cloudiness, up to 2050. They base on 39 models from the state of the art. In addition, they explain the influence these parameters will have on photovoltaic production. The model takes into account just RCP8.5, with 8.5 indicating the radiative forcing reached at the end of 21st century compared to preindustrial state. Clearly, this would depend on socio economical and policy decisions, but differences are not too large because divergence among models starts becoming important after 2050.

Hence, climate conditions are crucial parameters to analyse, in order to face non-dispatchability of renewable sources. Clearly, uncertainty affects them, forcing the use of different techniques to model them properly. Additionally, climate change intensifies the problem, because of difficulties in predicting data in an accurate way. Moreover, the "weight" of weather conditions has to be evaluated for each specific case. Indeed, a net zero energy systems will be strongly influenced by them, because the renewable production does, while a fossil fuel based system can be studied without their introduction.

1.2.2 Price of energy carriers

These parameters are fundamental to understand the economic advantage of a solution. Of course, as explained in the previous section, the necessity of introducing them is strictly related to the energy system itself, depending on the components.

However, price evolution is difficult to be forecasted with accuracy: because of this, many researchers assume them as constant, such as [14, 16-19].

Focusing on natural gas, many ways are used by researchers to introduce its uncertainty nature:

- In [5], as already explained, design is based on deterministic curves, thus uncertainty is not introduced. In a second step, natural gas price is varied through local sensitivity analysis, to see if its variation implies relevant variations in results.
- Authors in [6] use Swiss Energy Strategy 2050 [23] to get natural gas values. Unfortunately, only two scenarios are presented, implying a limited variation in future prices. To solve the problem, a "cone" of variability is used, increasing it year by year (±2% on first year, ±4% on second year, ±40% for last year). Finally, a uniform distribution is generated, in order to get the scenarios with an equal probability.
- Mavromatidis et al. [21] and Wang et al. [24] generate a uniform distribution with a set variance around a deterministic base value.

The previous discussion underlines two fundamental points when dealing with natural gas prices (analogue for biomass). First, the necessity of choosing reliable sources to get future scenarios, because of the strong uncertainty associated with the parameter. However, these sources can give just deterministic values, forcing to associate probability distribution functions. The second issue is a consequence of the first, because the probability associated with each scenario has to be calculated: in the state of art, uniform distributions are usually adopted, giving equal probability to each possible value.

To deal with the first problem, the only possible solution is to refer to reliable documents. If local forecasts are not available, a possibility is to use IEA World Energy Outlook 2022 [25] or [26] as references. They present three possible scenarios:

1. The Stated Policies Scenario (STEPS), taking into account only specific policies announced or in place by governments.
2. The Announced Pledges Case (APC), which assumes that "all announced national net zero pledges are achieved in full and on time, whether or not they are currently underpinned by specific policies".

3. The Net-Zero Emissions by 2050 Scenario (NZE). This shows how the global sector should act to reach decarbonisation in 2050. Note that this is *one* way to reach it, not *the only* way.

Assuming to work with this reference, some tricky points have to be marked:

1. Predictions show just final prices, for each scenario, for 2030 and 2050. Therefore, they do not show the evolution of them, avoiding any discussion about price volatility, which is too hard to be included.
2. Focusing on NZE, this scenario shows one possible solution for decarbonisation, which is not the only possible one, as well as it will not probably be the adopted one. Therefore, a set of possible scenarios, as done in [6], should be considered.
3. No information about scenarios probability is done. Clearly, the common solution is to give equal probability to each scenario, through a uniform distribution. Differently, an interesting decision would be to ask the customer which scenario is the most probable, modelling probabilities on that base. For instance, if he does not think decarbonisation is possible, probability of STEPS has to be increased, as well as the one for NZE must decrease.

Further problems are related to electricity prices. Solutions may be the ones found in the Literature:

- In [5], as already explained for natural gas, it is not treated as uncertain parameter in the design phase. In a second step, it is varied through sensitivity analysis
- In [6], the "cone" of variability is still used, with a lower variability with respect to natural gas (±1% instead of ±2% for the first year and so on).
- In [18], electricity price increases with a set inflation rate. This means that every year its price increases by 3%.
- In [21], a uniform distribution around a deterministic value is used.
- Wang et al. [24] use Gini index to convert the continuous variation to a discrete number of points with their probability.
- In [27], each day of the previous year is taken as a possible scenario. The first solution is obtained with equal probability for each scenario, while the second one through sensitivity analysis, thus increasing the probability of scenarios with higher volatility.

As for natural gas, many researchers [14, 16, 17, 19] consider electricity price as constant.

Differently from fossil fuels, electricity price is even more difficult to be forecasted. Indeed, it depends on the structure of the market, which is going to vary significantly in the future. Based on this consideration, Domìnguez et al. [7] present a multi-stage stochastic model to analyse the day-ahead market and the balancing one, once the configuration of the first model is set. Note that electricity price is not assumed, but is calculated, starting from the European energy production structure. Golombek et al. [28], to assess the role of transmission and energy storage at 2050, use LIBEMOD [29], a specific deterministic model, which is able to find prices by taking policies as inputs. These two papers are indicated to show how complicated models should be to get optimum forecast results.

In order to find a coherent way to assess future electricity price, Weron [30] reviews forecasting models, dividing them in three categories:

1. Short-term electricity price forecast (EPF), involving forecast from a few minutes up to a few days ahead, of primary importance for day-to-day market operations.
2. Medium-term horizons, from a few days to a few months ahead, generally preferable for balance sheet calculations, risk management and derivatives pricing.
3. Long-term EPF, concerning months, quarters or even years. Their goal is to evaluate investment profitability, analysis and planning, such as determining the future sites or fuel sources of a power plant. They usually deal with capacity-investment decisions, avoiding unit-commitment ones.

Design and operation problems normally deal with a long-term optimization period, usually matched with the system lifetime. Therefore, they should refer to long-term electricity price forecast, which are not indicated in the paper. However, some models could refer to shorter periods: for instance, the one to get a net present value (NPV) equal to zero. Additionally, to reduce computational effort, a choice could be to find

a solution for one year as operation, to test it in a second moment for a longer period (Section 4.4). Hence, examined methods are reported for these problems:

1. Multi-agent models (Nash-Cournot framework, supply function equilibrium, strategic production-cost, agent-based). Cost-based models are capable to forecast hour by hour and bus-by-bus, ignoring strategic bidding prices, while equilibrium approaches take into account them too. However, both of them work by matching demand and supply curves in the market. Author distinguishes among optimization models, which are not interesting for EPF because of their intrinsic goal, equilibrium models and simulation ones.
2. Fundamental models (parameter rich fundamental, parsimonious structural), trying to deal with physical and economic relationships present in the production and trading of electricity, such as loads, fuel prices, wind power and temperature.
3. Reduced-form models (jump-diffusions, Markov regime-switching), developed not to provide accurate hourly price forecast, but to give main characteristics of daily prices. They could easily become inaccurate or too complex to be solved.
4. Statistical (Similar-day, exponential smoothing, regression models, AR-type, ARX-type, threshold AR, GARCH-type), good for derivatives valuation and risk analysis, but not too precise for forecasting electricity prices. They are based on a combination of previous prices or current values of exogenous variables. For example, similar day consists in searching in history days with similar defined characteristics in order to compare their price trend.
5. Computational intelligence (Feed-forward neural networks, recurrent neural networks, fuzzy neural networks, support vector machines), combining elements of machine learning, evolution and fuzziness, wants to adapt to complex dynamic systems.

In any case, these are complex models, created to adapt to specific market conditions. For example, [31] refers to Nordic market, developing a framework difficult to use in another context. Furthermore, the creation of such models would require a big effort, impossible to match with the topic of the thesis. Precision is not guaranteed, because, working on twenty years long periods, possible mistakes will propagate, entailing unfeasible solutions.

Gabrielli et al. [32] specifically focus on long-term electricity price forecast. First, they distinguish between:

- Data-driven models, which are very accurate about hourly prices. Unfortunately, they lack in long-term precision because of price volatility.
- Market-based models, whose goal is to comprehend the underlying mechanisms of the electricity market. They are not good for hour-to-hour results and they are not easy to adapt to markets different from the initial one.

To analyse price evolution, authors decide to adopt a data-driven model based on Fourier analysis. The price is thus descripted by its main characteristics (main frequencies) and fluctuations (other frequencies), while Fourier coefficients are calculated from historical and predicted values of several price drivers. To account out-of-sample issues, the model is coupled with a market-based one.

Results show that this model is able to follow prices in the UK market with a mean absolute percentage error equal to 10% on yearly basis. Therefore, to use such a complex model is not useful for this work, because effort is not justified by accuracy.

Therefore, prices of energy carriers are fundamental parameters to be modelled, because their values strongly influence the solution of the problem. However, their evaluation is complicated, because of the uncertainty in future values. Regarding fossil fuels, reliable future scenarios are required to obtain deterministic values, even if results are consequently linked to these assumptions. Furthermore, there is a lack of variability in results, as well as probability associated with each scenario is arbitrary. Indeed, uniform distributions can be adopted, while other solutions are associated with choices of the researcher. Instead, electricity price is over-technical, because its evolution depends on aspects as market structure. If the model considers a big system, such as a nation, electricity price can be calculated precisely. Alternatively, in the Literature forecasting models are presented.

Another solution would be to set a value of energy carriers for a simulation and then to vary it, like sensitivity analysis, to see how design choices are linked to them. In particular, it could help finding solid technologies, used in many scenarios. Note that this procedure is different from the one indicated in [5], because these parameters cannot be varied just after the design, whose convenience strongly depends on their values.

1.2.3 Energy demand

Energy demand includes electricity, thermal and cooling demand. Their evaluation is crucial, because satisfaction of loads is usually the main requirement. However, some aspects make it complicated:

1. It is hard to forecast energy demand, considering the "quantity". Indeed, according to the presented scenario, it could increase or decrease. For instance, in [25] STEPS and APC scenarios bring to an increase in demand, whereas one of the NZE key topics is to avoid it. Additionally, the trend for industrial or civil users may be different.
2. Even harder is the "structure" of the energy demand. As already indicated, electric, thermal and cooling demand are taken into account. Assuming to work with a civil user, according to APC scenario demand will increase. In any case, electrification influences it, thus electric demand is going to increase, while thermal demand is going to decrease because of substitution of boilers with heat pumps. Technical performances will influence it as well.
3. Possibility of demand side management (DSM). Zheng et al. [11] developed a model for the design and optimization of a microgrid focusing on this aspect. They distinguish between dispatchable (or capable of being shifter) and non-dispatchable loads, entailing a flexibility in the system management. Then, they use the sliding time window method for DSM [33].
4. Occupants' behaviour. This aspect is usually neglected, but it could influence analysis, especially nowadays, where users start being more responsible.

In the Literature, many ways are used to consider uncertainty for demand curves:

- In [6], a Monte Carlo Building Performance Simulation (BPS) is adopted. First, a probability description is associated with input of the BPS. Then, several profiles are generated to have different scenarios.
- In [7], authors generate three scenarios by following IEA predictions about increasing in demand.
- In [12], demand side management is performed. Therefore, starting from deterministic curves, these are varied hour by hour following the method described in the article.
- In [15], simulations in Simulink are done with respect to 40 different buildings with their own characteristics, to collect then results in a one-year curve.
- In [16], they collect historical data and normalize them. Then, they distinguish between winter and summer case to proceed with a k-means clustering process. Probabilities are evaluated as cluster frequencies.
- In [17], a uniform distribution around a deterministic profile is used to describe electricity demand.
- In [34], bootstrap simulation is used to get demand distribution.

In many works, such as [5, 14, 15] demand curves are assumed as deterministic ones.

Following previous examples, there are different ways to model uncertainty regarding demand curves. According to Mavromatidis et al. [6], however, probability distribution functions are inadequate to do that. Indeed, this does not take into account autocorrelation of these time series, as well as relations with other parameters such as climate conditions or occupant behaviours.

1.2.4 Interactions among uncertainty elements

Most of the papers in the Literature assume each uncertainty parameter as independent from each other. For example, in [12], authors do it using probability distribution functions. Unless only one uncertain parameter is taken into account, this solution is not accurate, as explained in previous section [6]. Indeed, some relations occur between parameters. For instance, if the temperature increases, heat requirement is going to decrease, as well as cooling requirement is going to increase. Another example is the relation between electricity and natural gas prices, which can be very complex to evaluate.

However, some papers try to correlate uncertainty parameters. In [5], building performance simulation is repeated many times, coupled with Monte Carlo method, to correlate energy demand and renewable production. In [35], a time-varying parameter vector autoregressive model with stochastic volatility model is implemented to study dynamic relationships among oil price, renewable energy consumption and climate policy uncertainty (CPU). This last parameter is represented by the respective index introduced in [36], based on "the number of articles including inherent topics". Results show that the increase in CPU causes an increase in the oil price in the short and medium term, while it tends to converge in a long term: this is in contrast with other studies mentioned by the authors, resulting in a low reliability. They continue explaining that, in general, impact of CPU increase is positive on short and long term to renewable energy consumption, but negative on the medium one, even though the answer of five different types of renewable sources is "heterogeneous".
These last papers [35, 36] are reported to underline both the difficulties in relating parameters and finding reliable results. In fact, the risk is to force the work, as done in [35], basing on an ambiguous index and finding results in contrast with other sources.

Olanipekun [37] et al. investigate the relation between the global use of renewable energy sources and oil price uncertainty, by the use of advanced econometric techniques (Wavelet Coherence and Quantile-On-Quantile Regression). Results prove that an increase in the use of renewable energy entails a lower oil price uncertainty and that the impact of renewable energy on geopolitical oil price uncertainty is greater than the opposite influence. This is reported to show how complex could be to relate such parameters, which is out of scope for this work and, furthermore, is out of scope for optimization problems. Please note that data reduction is required to reduce complexity and these models would require an incredible computational effort to be implemented.

As already explained, the goal is not to find a perfect relation among uncertainty sources, but to analyse their influence on design phase. Hence, if many uncertain parameters are introduced, flexibility, defined as "the capacity of a system to resist performance degradation due to changes in the external environment", could be evaluated to understand the quality of the results. Perera et al. [16] propose a procedure to evaluate it.

1. Formulation of scenarios for the stochastic optimization.
2. Values of indicators are evaluated for each scenario based on the time series simulation.
3. Degradation, defined as the variation of a parameter with respect to a reference value, which can be the expected value of the indicator, is calculated.
4. Some ranges are introduced in order to normalize the different parameters, obtaining an S*P matrix (S=number of scenarios, P=criteria). Clearly, a set variation can give a different indication depending on the parameter. For example, a decrease of 15% of renewable generation could be accepted, while an increase of 15% of the NPV is more serious.
5. A weight is introduced for each parameter in order to underline the importance it has. Finally,

$$\mathbb{E}(IPD) = \sum_{s \in S} weight_s \sum_{p \in P} parameter$$

$$Flexibility = \frac{1}{1 + \mathbb{E}(IPD)} \qquad (1.5)$$

1.2.5 Clustering techniques for data series reduction

Computational optimization for the design and operation of energy systems is not a trivial issue. Indeed, the state of the art include complex mixed linear or non-linear programming models: clearly, the solution has to be as precise as possible, without entailing an unaffordable computational time to get it. Time-series aggregation is thus a fundamental topic, whose relevance is proved by its use in the Literature.

Clustering techniques allow reducing computational effort. The idea is to obtain representative periods by aggregating similar elements. A perfect clustering process should reduce the computational time by orders of magnitude while representing the whole information of the original dataset. However, this is impossible, because data reduction implies persistent or random errors [38].

As suggested by [38], some passages are required for the whole process. Let us consider a dataset with I rows, corresponding to the candidates as representative elements ($i \in \{0, \dots, I\}$) and J*A columns, with J corresponding to the daily time steps for each i element ($j \in \{0, \dots, J\}$) and A indicating the number of attributes ($a \in \{0, \dots, A\}$). For example, collecting one year of data for solar irradiance, wind speed and air temperature, with hourly resolution, the matrix will have I=365 rows and J*A=24*3=72 columns, as shown in figure 1.2.1.

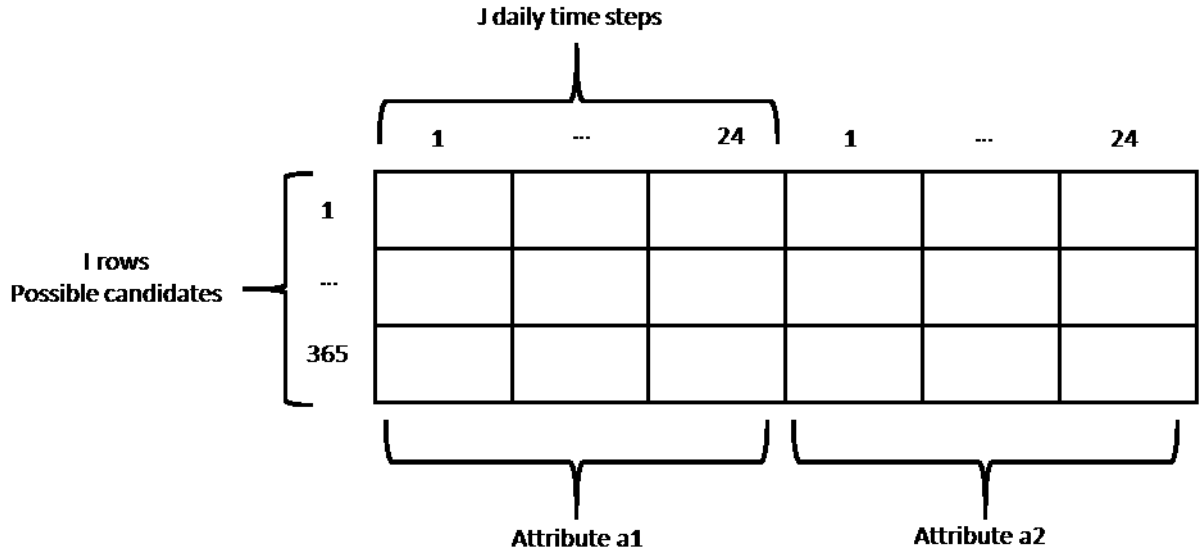

Figure 1.2.1: Organization of dataset for data series reduction.

The first step is data normalization. In particular, this is crucial if there are elements of different nature in the initial dataset. For instance, if global solar irradiance [W/m^2] and air temperature [°C] are considered, differences in their magnitude do not allow a proper clustering process. Normalization can be referred to the maximum element, bringing back each data to [0,1] interval, or a z-normalization, with a mean value equal to 0 and a standard deviation equal to 1. Moreover, three types of process are possible, assuming a 0-1 normalization: full normalization, comparing all elements to the maximum one, which is not good if the number of attributes is higher than 1; element-based normalization, working on each daily time step; sequence-based normalization, working on single days.

For instance, assuming to work with a dataset with only global solar irradiance, with 365 candidate days and 24 hours as daily time step, full normalization will assign 1 to the maximum value of irradiance in the whole year, scaling the others with respect to that. On the contrary, element-based normalization scales with respect to the maximum element of each column, thus the maximum value of each hour: this procedure is usually adopted for k-means, k-medoids or hierarchical [38]. Last, sequence-based normalization works day per day: this is used for k-shape method or Dynamic time warping barycentre averaging (DBA) [38].

Kotzur et al. [39] suggest the adopted technique for this work, working with a 0-1 element based normalization

$$x'_{a,i,j} = \frac{x_{a,i,j} - min\, x_{a,j}}{max\, x_{a,j} - min\, x_{a,j}} \quad (1.6)$$

Second aspect is related to the considered distance measure. In this thesis, Euclidean distance (ED), which is the most common for "classic" techniques as k-means, k-medoids or hierarchical, is used

$$dist(X,Y) = ED(X,Y) = \sqrt{\sum_{j=0}^{J} x_j - y_j} \quad (1.7)$$

Therefore, the distance is evaluated for each daily time step for the respective candidates. Clearly, the following equation considers just J columns because the assumption is to work with A=1, thus only one attribute.

Alternatively, instead of using ED, Dynamic Time Warping (DTW) is an option, usually adopted with DBA techniques. ED is a particular case of DTW with a warping window equal to zero. More information can be found in [40]. Last, shape based distance (SBD) is another distance measure introduced by Paparrizos and Gravano [41].

As already explained, many clustering algorithms are possible. Averaging periods is a possibility, as done by [42]. However, as proved by [39], this method is the worst one in terms of representing the variability of the dataset. Hierarchical clustering method allows obtaining representative periods by merging in the same cluster the two closest elements (in terms of distance) for each iteration. Applications can be found in [43] and [44].

k-means generates the clusters to minimize the squared error between the mean of a cluster and all possible candidates. The problem can be defined as

$$min \sum_{k=1}^{K} \sum_{i=1}^{I} \left[\sum_{j=1}^{J} \sum_{a=1}^{A} \left(x_{a,i,j} - \mu_{a,j,k} \right)^2 \right] \times \delta_{i,k} \quad (1.8)$$

In the previous expression, $\mu_{a,j,k}$ indicates the centroid of the cluster k for daily time step j and attribute a. $x_{a,i,j}$ is the element of the candidate day i, time step j and attribute a. $\delta_{i,k}$ is a binary variable equal to 1 if the candidate belongs to the cluster k, 0 otherwise. To be sure that every candidate belongs to a cluster, equation (1.9) is required

$$\sum_{k=1}^{K} \delta_{i,k} = 1 \quad \forall i \quad (1.9)$$

The problem is a mixed integer non-linear problem (MINLP), difficult to solve. Therefore, a simplified algorithm is adopted, which converges by the way to a local minimum.

1. The number of clusters is decided a priori. Randomly select an initial partition of clusters.
2. Assign each element to the closest cluster.
3. Calculate the new centroid.
4. Repeat until convergence.

More information about theory of k-means can be found in [45]. Adhau et al. [46] use k-means to study availability of a micro hydro power plant, Fazlollahi et al. [47] use it studying multi-objective optimization of distributed energy systems. [17, 18] use k-means to generate scenario for two-stage stochastic problems.

The choice of the representative period for each cluster is interesting. Indeed, instead of the centroid, the medoid or the closest element to the centroid can be used too [48]. Centroids are calculated by averaging terms: this reduces peak conditions, which are useful for design of energy systems.

k-medoids is similar to k-means, but the representative element is the medoid instead of the centroid. As explained in [49], the problem is stated as a MILP, thus a global optimum can be found. However, a high computational effort is required, so the algorithm will not be used for this work. However, Domínguez-Munoz et al. [50] adopt it to find typical days for demand, while Mavromatidis [21] uses it to reduce scenarios for the two-stage stochastic optimization dealing with distributed energy systems. Graphical comparison between k-means and k-medoids is presented in figure 1.2.2.

Other methods, such as k-shape [41] or Dynamic time warping barycenter averaging (DBA) clustering technique [40] are possible, but out of scope for the paper. The best clustering technique does not exist: depending on the application, one of them will perform better and it is difficult to predict which one before trying it.

Another crucial aspect is the addition of extreme scenarios, such as peak of thermal demand or electricity price. In fact, these elements are not representative of a period, so they are far both from centroids or medoids. Nevertheless, they are very important for the design of energy systems, because energy availability could not be guaranteed without them. Some alternatives are possible for this issue:

1. No introduction of extreme periods.
2. Adding these periods as representative ones, in addition to the ones obtained through the clustering process. Note that, however, the difficulty here is the calculation of weight (or probability) for these new clusters.
3. Forcing them to be new clusters centres, repeating the clustering process for another iteration, to re-assign candidates. Note that the influence will be low, because only few candidates will be assigned to these new clusters, implying a lower weight (or probability).
4. The extreme periods become the new cluster centres for the clusters they are assigned to. This solution usually overestimates their weight on the total problem.

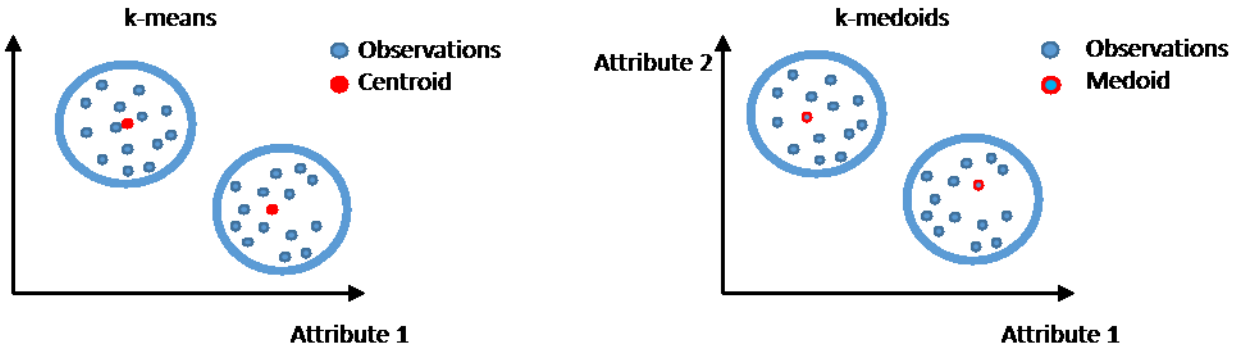

Figure 1.2.2: Difference between k-means (left) and k-medoids (right). The centroid is calculated as the average value and it may not exist, while the medoid is a real element of the cluster.

In this thesis, k-means is used to obtain representative days (Section 4.3) and scenarios (Section 4.4). Nevertheless, because of the obtained average profiles, the solution could not be too precise. Hence, instead of the centroid the closest element to it is taken as representative element for the cluster. k-medoids is not adopted because of the unaffordable computational effort, due to the heavy dataset. Actually, according to [38] k-means tends to underestimate the objective function, reducing the error while increasing the number of clusters, while k-medoids goes closer to the real value, but represents worse the entire dataset, because the medoid is taken instead of the centroid. The introduction of extreme scenarios, by the way, changes a little these trends.

Another point is the number of clusters. As already explained, for k-means it has to be decided a priori, which means there is no sureness about the quality of the chosen number. Normally, heuristic methods are used and quality of the optimized solution is not proportional to the quality of clusters.

In [38] authors use the within-cluster sum of square distance to evaluate the quality of the clustering process. Related to that, they find that it increases, with k-means, if the number of clusters increases. Anyway, this is not actually true. Indeed, if number of clusters increases, the distance of the elements belonging to a single cluster surely decreases, but there the possibility of having bad-positioned elements increases as well.

A coherent way to evaluate the quality of clusters is to create a profile of Average Silhouette Value (ASV) and keep the number of clusters equal to the one that maximizes this parameter. According to Roussew [51], the silhouette value for a single element can be defined as

$$sil(i) = \frac{a_{sim}(i) - b_{dis}(i)}{max(a_{sim}(i) - b_{sim}(i))} \qquad (1.10)$$

In which $a_{sim}(i)$ is the average similarity the element has with respect to the all other elements in the cluster it belongs, while $b_{dis}(i)$ is the average similarity with respect to the most similar cluster it does not belong to. The silhouette value, for cluster k, is given by the average of the value of sil(i). It seems obvius that

$$-1 \leq sil(i) \leq 1$$

Therefore, a coherent solution is to choose the number of clusters K that maximizes the ASV. In fact, if K is too low, the ASV will be low, because $a_{sim}(i)$ cannot be high, due to differences of elements in the same cluster. On the contrary, if the number of clusters is too high, ASV is low, because $a_{sim}(i)$-$b_{dis}(i)$ will be, due to the fact that an element can stay both in cluster A or B. Note that, for each element, a value of ASV equal to zero means it could stay in another cluster as well, while a negative value means it is bad-assigned.

Returning to [38], the highest SSD does not correspond to the highest quality for clusters and does not correspond as a good quality for the solution too, as proved in this thesis. The best solution should be to choose the number of clusters equal to the one corresponding to the "elbow" of the curve representing the SSD.

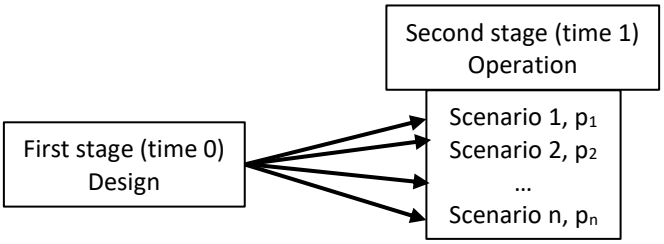

Figure 4.3.1: Scheme of the two-stage stochastic problem

1.2.6 Theory of stochastic programming

In this work, the adopted method for modelling uncertainties is stochastic programming. The general formulation of a two-stage stochastic problem, as stated by Infanger [52], is

$$\min_{x,y} c^T x + \sum_{s \in S} v_s^T y_s$$

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

$$-B_s x + D_s y_s = d_s$$

$$x, y_s \geq 0 \tag{1.11}$$

Where c are the first-stage coefficients, v the second-stage ones. A and b are coefficients related to the first-stage constraints, while D and d are the ones related to the second-stage. B is called transition matrix and couples the two different stages.

The idea is that at time 0 a decision, whose convenience depends on the outcome at time 1, has to be made. On an ideal point of view, the best choice is to wait for time 1 to decide, in order to see which scenario occurs: this is the *wait and see* approach. Obviously, it is not implementable in practice, but it allows understanding how good the proposed stochastic solution is. For each scenario s, an optimal solution z_s, which is a lower bound of the objective function, is found.

$$z_s = \min_x f(x, s) \tag{1.12}$$

Unfortunately, in a real situation the decision must be done at time 0, before knowing which possible scenario realizes. This is the *here and now* approach, which bases on the minimization of the expected value of the objective function.

$$z_{HN} = \min_x \mathbb{E} f(x, s) \tag{1.13}$$

This approach is applicable to a design/operation problem, because design decisions, affected by what happens in the operation phase, must be done before the operation starts. For instance, assuming to install a PV system at time 0, its convenience depends on the global solar irradiance for its whole lifetime: if weather is rainy for the following twenty years, there is no convenience in installing it; on the contrary, with only sunny days the chosen size should be high. Hence, the ideal decision should be to wait for the following days (or years) and see what happens. Unfortunately, this is not possible, forcing to make a decision here and now, basing on the expected value of the uncertain parameter.

Note that the number of stages is not a trivial aspect. Indeed, if it is equal to two, it means the first stage is for the design (time 0), whereas all the operation is included in the second stage (time 1). Hence, if the optimization period is one year, everything is included into a single stage, which means there is not possibility of correcting it throughout the period. The alternative is to use more than two stages, which is by the way interesting only if there is a change in information. For instance, Dominguez et al. [7] consider more stages because design is done many times, to substitute or add components. Another example is the electricity market, which should be a three stage stochastic problem [53], with the first stage for the design, the second stage for the day ahead market and the third one for the balancing market.

Nevertheless, most of works in literature deal with just two stages. For example, Gomes et al. [54] develop a model for management and operation planning of a microgrid, modelling it as a two-stage stochastic problem. Similarly, De et al. [55] focus on strategic bidding for a generation company, always with a two-stage stochastic model. Ghaemi et al. [13] work to perform two-stage stochastic programming in a district

energy system. In particular, uncertain parameters are solar irradiance, wind speed, electricity demand and CO_2 emissions. They model them correcting the collected data, generating scenarios, reducing them to limit computational effort and, finally, preserving important scenarios, as extreme electrical requirements day is. Zheng et al. [18] develop a multi-year model for the design and operation of residential photovoltaic-battery systems. This is framed as a two-stage stochastic problem, with energy demand and solar irradiance as uncertain parameters. Scenarios are created with k-means clustering technique. Golombek et al. [28], to see the role of transmission and energy storage in Europe at 2050, build a model with investment at 2020, 2030, 2040 and 2050, implying a multi-stage stochastic problem.

In [18] authors develop a model for the convenience of a residential battery-PV (photovoltaic) system under uncertainty. They use stochastic programming to solve the problem, underlying the advantages compared to a deterministic solution. In [19] authors try to compare deterministic and stochastic solutions working on multi-objective optimization for a distributed energy system under uncertainty, finding that stochastic solutions are more robust.

However, according to Zhou et al. [56] stochastic programming does not entail significant improvements, so deterministic model for the design is preferred due to the computational efficiency. As indicated by Mavromatidis [21], stochastic programming requires huge computational efforts, because of the number of scenarios considered (in that case, 1215), compared to a deterministic model, whose solution can be found with a common laptop. Most of researchers, indeed, still work on design of energy systems with deterministic models. For instance, [57] assesses a method to evaluate long-time uncertainties for a photovoltaic system: even if the model considers relations among external factors and internal factors are included as well, it is a deterministic one.

Clearly, other methods are possible and can be found in the Literature. For instance, robust optimization is a possibility. Reference [58] is a paper about optimization of an energy hub including uncertainties, which are renewable energy production, multi-load demands and electricity/gas prices. Authors model the problem as a two-stage robust optimization one, with the first stage for the design and the second one for the operation. The formulation of the objective function is thus (1.14):

$$min\, Cost_{DES} + max(min\, Cost_{OP}) \tag{1.14}$$

With $Cost_{DES}$ design cost and $Cost_{OP}$ the operation one. The idea is to find an upper bound for the operation cost, to be sure the every possible outcome entails a lower cost solution. Clearly, precision is required not to overestimate that.

To solve an optimization problem, the objective function is one of the first things to decide. In particular, it could focus on economic aspects, such as the minimization of the total cost, environmental aspects, such as minimization of CO2 emissions, as well as social ones, such as customers' satisfaction. Clearly, if in the first case the cost is easy to be evaluated, in the other ones some parameters should be introduced to relate the use of a component to its impact on the function. In the Literature, some works, such as Da Lima et al. [17] or Liu et al. [59], deal with multi-objective optimization, creating Pareto frontiers to find the best solution regarding total costs and emissions. However, other works, as [21], just consider economic aspect, while environmental issues are included through a constraint that limits CO_2 emissions.

1.3 Goals and contributions

The general goal of this work is to study the design and operation phases of an energy system under uncertainty. Results should explain whether the modelling of uncertainty associated with some parameters helps improving design choices, such as components sizes. Specifically, the uncertain parameter under consideration are the global solar irradiance and the air temperature.

To understand that, a residential multi-energy system (MES) is taken into consideration. The idea is to be in 2010, with the necessity of satisfying electrical and thermal requirements: historical data about global solar irradiance and air temperature are available just for the past, specifically for the period 2005-2009, but convenience in design choices depends on the future. Therefore, deterministic (Section 4.3) and stochastic models (Section 4.4) of such system are developed, to compare design solutions and total costs to the ones obtained with a perfect model for the "future" (Section 4.1), hence for the period 2010-2020.

The scheme for the residential MES is presented in Section 2; input data, except for the uncertain parameters, are introduced in Section 3; Section 4 describes the different models; results are presented and discussed in Section 5; Section 6 concludes the work.

The following points are deepened throughout the thesis:

1. Critical analysis about the introduction of air temperature as attribute for the clustering process. Indeed, in the Literature global solar irradiance is rarely coupled with this parameter, whose relevance for the representation of the initial dataset is strong (Section 5.1).
2. Comparison among different methods to generate representative days for a residential MES, to see which is the most suitable to use for the design phase of an energy system. In particular, in the state of art different methods are used, as indicated in the previous Sections, usually without explaining the advantages. Clustering techniques are compared with average seasonal and monthly profiles. Generation of seasonal clusters is discussed as well.
3. Introduction of an innovative two-step clustering process to generate scenarios for representative days. In fact, some works consider deterministic models as stochastic ones, just generating typical days for few uncertain parameters. The idea is to generate scenarios for each representative day, to see if it could help obtaining precise solutions.

2. Object: Residential multi-energy system (MES) scheme

The goal of this work is to study the design and operation phases of an energy system under uncertainty. In particular, results should explain whether the modelling of uncertainty associated with global solar irradiance and air temperature helps improving design choices, such as sizes of the energy conversion units.

To answer to the previous question, a residential multi-energy (MES) system is taken into consideration. Deterministic models, with N representative days as optimization period, and two-stage stochastic model, with N representative days and M different scenarios for each day, of this system are developed, in order to understand which method allows obtaining "better" solutions, thus closer to the reference one (Section 4.1).

In particular, the assumption is to be in 2010, with an electric and a thermal load to satisfy: the aim is to minimize the total cost for the users for the entire period 2010-2020, which is the "future". Indeed, historical data about global solar irradiance and air temperature, the uncertain parameters, are available just for the past, specifically for the period 2005-2009. Therefore, historical data must be used to forecast the future, to discuss the convenience in installing components, their sizes and to optimize the operation. In order to test the quality of the given solutions, obtained solving the different models, a comparison is done with a reference model for the period 2010-2020, which has perfect knowledge about the "future".

Therefore, the period of study is divided in training dataset, for the period 2005-2009, and testing dataset, for the period 2010-2020. The training dataset allows doing an "in-sample analysis", because scenarios are obtained through clustering techniques from its data; testing dataset, on the contrary, entails an "out-sample analysis", testing the solution on a larger set of scenarios.

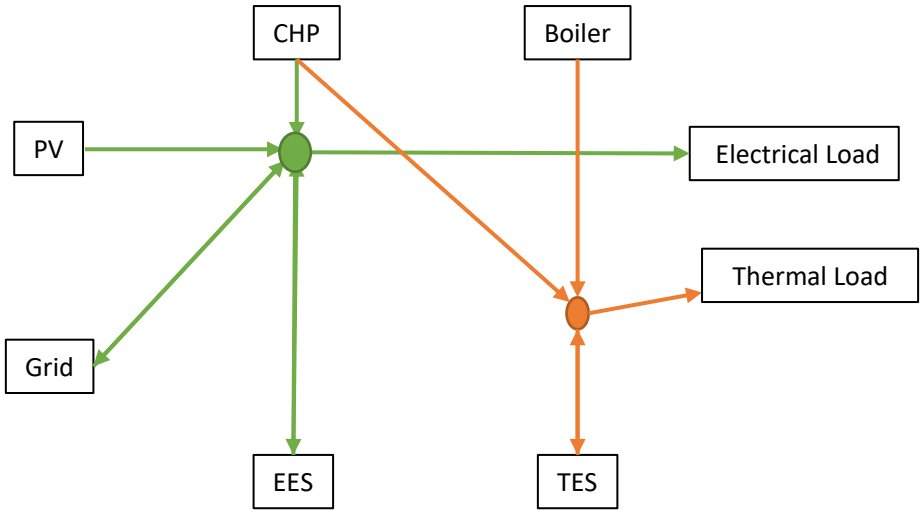

Figure 2.1: Schematic energy system scheme. The electric side consists in a photovoltaic system (PV), an electric storage (EES), the national grid and an electric load. The cogeneration system (CHP) is included both in the electric and thermal side, which considers a boiler, a thermal energy storage (TES) and a thermal load. If a component can give energy to the system or take energy from it, a bidirectional arrow is connected to it.

Figure 2.1 shows the scheme under analysis. Suppose to have a group of residential loads, which have to be satisfied hour per hour, both on the electric and thermal side. Note that, because of the hourly resolution, concepts of energy and power coincide. They would spend in any case a given amount of money for their energy requirements, so the study focuses on which energy conversion units allow them to minimize the total cost for that time period. This cost [k€] is seen as sum between the design and the operation one.

Furthermore, the auto-consumption of energy is not considered a revenue, while the excess in electric energy sold to the grid is seen as a negative cost for the objective function, presented in Section 4.

To fulfil load requirements, the following components are taken into account.
On the electric side components are:

- A photovoltaic system (PV) [60], producing energy according to equations shown in Section 4.2.1.
- An electric energy storage (EES) analysed in Section 4.2.2.
- A cogeneration (CHP) system, with reference to catalogue [61], Section 4.2.3.
- The national grid, purchasing energy to the system or buying from it. Purchasing and selling prices are indicated in Section 3.2, while equations in Section 4.2.6.

On the thermal side:

- The CHP system, as already indicated.
- A natural gas fuelled boiler, with respect to reference [62], presented in Section 4.2.4.
- A thermal energy storage (TES), with respect to reference [62]. Equations in Section 4.2.5.

Techno-economical characteristics for these components are introduced in Section 3.

The system is supposed to be installed in Padova, Italy (45,406°N, 11,877°E). The models, explained in Section 4, have to evaluate the convenience in installing such components and, if so, their size. The number of the energy conversion systems can be zero, if they are not included, or one. Hence, two cogeneration systems, for instance, cannot be used.

3. Techno-economic data

This section includes information about deterministic input data for the models, excluding air temperature and global solar irradiance, which are the uncertain parameters. In particular, Section 3.1 considers characteristics of the involved technologies, such as efficiencies or prices; Section 3.2 considers profiles of energy carriers, thus electricity and natural gas; Section 3.3 presents energy demand profiles for this group of residential users.

3.1 Technologies characteristics

Technology	*Quantity*	*Unit*	*2020*
PV	InvestmentCost_var	€/kWp	1250
	InvestmentCost_fix	€	0
	Space Requirement	m2/kWp	9,7
	efficiency Standard	-	0,136
	efficiencyBalanceofsystem	-	0,85
	NOCT	°C	43,2
	b0	1/°C	0,454/100
CHP	InvestmentCost_var	€/kWel	1738,26
	InvestmentCost_fix	€	32046
	MinLoad	%MaxLoad	0,3
	F(P)_var	.	0,2722
	F(P)_fix	.	3,7125
	Q(P)_var	.	1,7246
	Q(P)_fix	.	9,3109
Boiler	InvestmentCost_var	€/kWth	64,86
	InvestmentCost_fix	€	1622
	MinLoad	%MaxLoad	0,3
	EfficiencyTh	.	0,97
EES	InvestmentCost_var	€/kWh	880,28
	InvestmentCost_fix	€	3494,44
	η_{EES}	.	0,87
	Self Discharge	%Size/hour	0,04
	Output Capacity	kW/kWh	0,3
	Input Capacity	kW/kWh	0,3
	SOC_{min}	%Size	0
TES	InvestmentCost_var	€/kWh	244,05
	InvestmentCost_fix	€	968,52
	η_{tES}	.	0,7
	Self Discharge	%Size/hour	2,1
	Output Capacity	kW/kWh	0,7
	Input Capacity	kW/kWh	0,7

Table 3.1.1: technical and economic characteristics of each energy conversion unit.

Table 3.1.1 shows the characteristics of the elements of the energy system. Economical parameters refer to a fixed and a variable costs, respectively c_{fixed} and c_{var}. The general expression for the design cost for a component, deepened in Section 4, is $Cost_{des} = c_{fixed} * \delta_{inv} + c_{var} * C_{component}$, where δ_{inv} is a binary variable describing the presence of a given unit and $C_{component}$ is the size of the unit. Then, technical

parameters, such as efficiencies or self-discharge capacities, represent each technology. This information is taken from references [62] and [63]. The assumption is to have O&M costs equal to zero.

Since the design is done at time 0, it does not have sense the consideration of evolution in energy components prices or performances, as explained in Section 1.1. Therefore, these parameters are deterministic, so nu uncertainty is considered for them.

3.2 Energy carriers

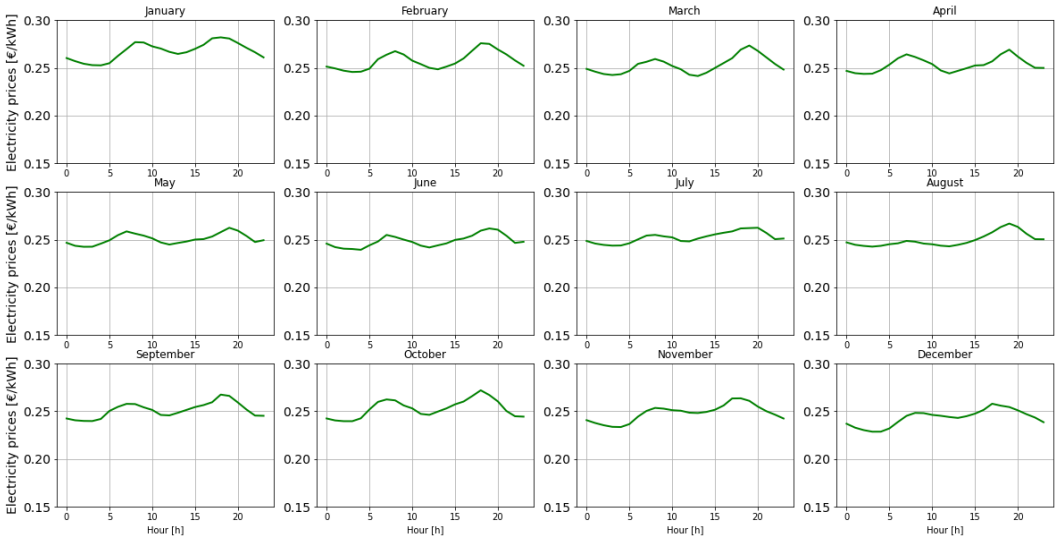

Figure 3.2.1: Average monthly electricity purchasing price [€/kWh].

Figure 3.2.1 indicates the average price of the electricity purchased by the system for each month. It varies hour by hour according to PUN [64], which means 365*24=8760 values are the historical data and can be used as input for the models. On the contrary, the selling electricity price is set equal to 4c€/kWh, constant all along the optimization period.

Natural gas price, required by the boiler and by the CHP system, is set too, with a value equal to 0,480€/sm^3.

3.3 Demand curves

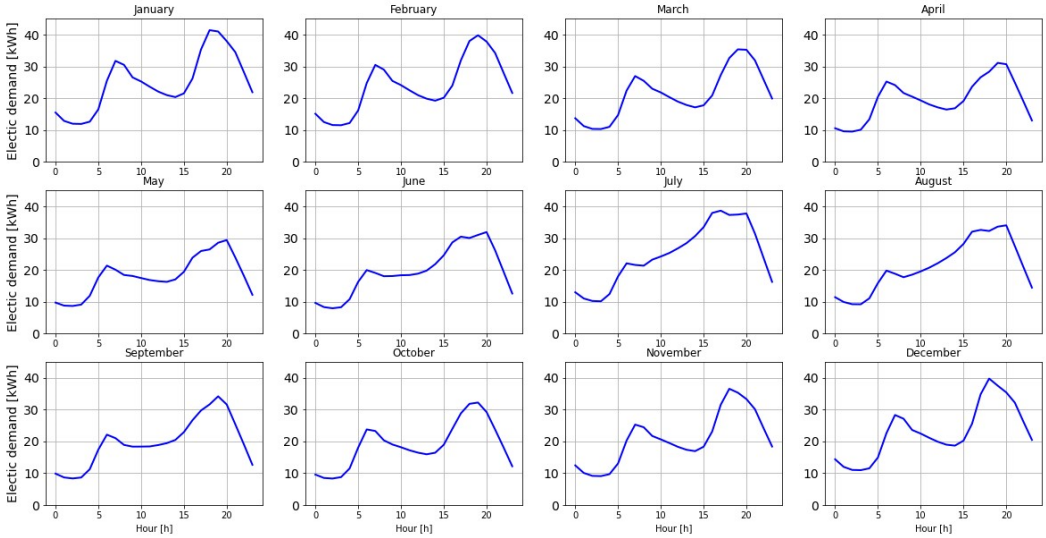

Figure 3.3.1: Average monthly electric load profiles [kWh] for the group of residential buildings.

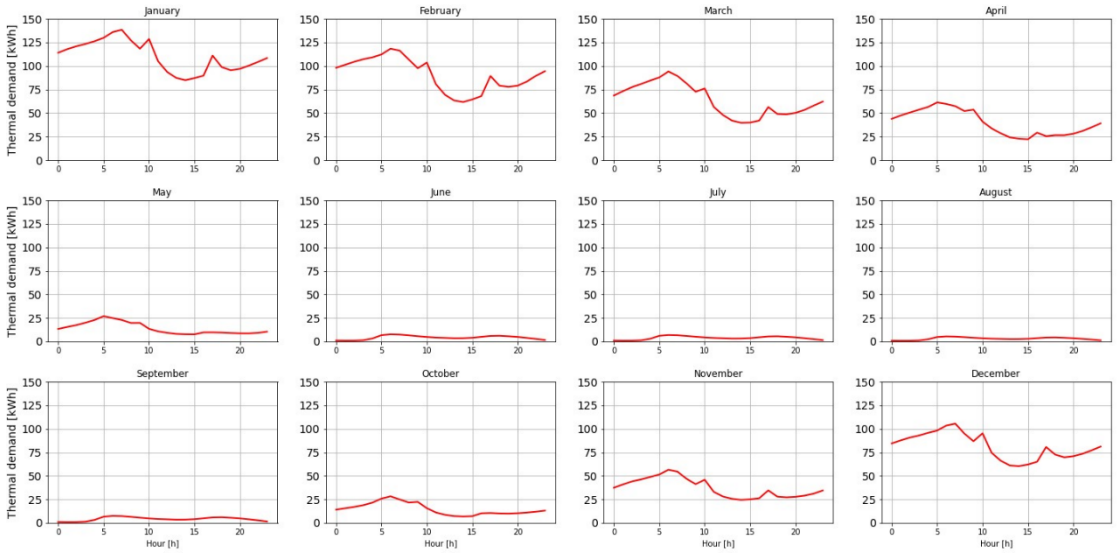

Figure 3.3.2: Average monthly thermal load profiles [kWh] for the group of residential buildings. Please note that it strongly depends on the season.

Figures 3.3.1 and 3.3.2 show the average monthly electric and thermal load for the group of civil users. Specifically, these values vary hour per hour, implying arrays of 24 elements per day.

As for electricity prices, demand profiles are taken from an internal database. Data refers to one single year, with resolution of one hour, entailing 365*24=8760 elements. Note that thermal load strongly depends on the considered month: this aspect will be crucial for the generation of representative days with clustering techniques, in Section 4.

4. Methods: deterministic and two-stage stochastic models of the residential MES

The following Section focuses on models of the residential MES. The goal is to describe input data, with special attention to the uncertain parameters, equations for the energy conversion systems, decision variables and the objective function. Section 4.1 introduces the reference model for the period 2010-2020; Section 4.2 explains in detail equations and inequalities for the components of the energy system; Section 4.3 describes the deterministic model with N representative days, whose generation procedures are explained, as optimization period; Section 4.4 analyses the two-stage stochastic model with N representative days and M scenarios; Section 4.5 is about robustness test, to understand the quality of a given solution.

4.1 Deterministic model of the residential MES for the period 2010-2020: decision variables and objective function

The following is a model describing the residential multi-energy system of Section 2. The assumption is to be at year 2010, with perfect knowledge about the future (period 2010-2020). Therefore, this model is the one that solves precisely the problem for this period, finding the best solution, according to the objective function.

In order to describe properly a model, focus should be on input data, equations/inequalities included and output data. Regarding this reference model, input data is:

- air temperature;
- global solar irradiance;
- electricity price;
- natural gas price;
- electricity demand;
- thermal demand;
- techno-economic data for energy conversion units.

Last three groups of parameters are described in Section 3. On the contrary, air temperature and global solar irradiance, which are the uncertain parameters under analysis, have to be described separately.

The global solar irradiance and air temperature, assuming perfect knowledge about “future”, are taken from [65], with an hourly resolution, implying a superimposition for the concepts of energy and power. In particular, this source gives values of these parameters for each hour of each day for the period 2010-2020. The 29th of February is excluded. Hence, 93690 values are taken into account and, due to the hourly resolution, the number of time steps is the same. However, this would be too heavy on a computational point of view, forcing to use clustering techniques to generate a set of representative days.

Let us focus on the clustering technique. In order to generate properly the representative days, k-medoids [55] should be applied. One of the differences between this method and k-means, as explained in Section 1.2.5, is the representative element for the clusters: for k-medoids, it is the medoid, while for k-means the centroid. Hence, k-means may not evaluate a real element of the dataset as representative one of the cluster, while k-medoids does. This is better, because it takes into account actual profiles of irradiance and temperature and, even more, it is possible to consider the date of these profiles, in order to couple them with their real electricity price and demand curves (data for one single year).

In order to generate clusters from a dataset in Python, TSAM [66] or sklearn [67] packs could be used:

- TSAM (Time Series Aggregation Module) generates clusters with k-means, k-medoids, k-maxoids, hierarchical or adjacent periods. It is very simple to use and it allows considering different physical quantities in the same process, even introducing extreme condition scenarios. However, some information, such as which elements are included in a cluster, are lost. Furthermore, the huge problem is due to the computational burdens in using k-medoids, which is the biggest limit in using TSAM. In fact, a dataset with 93690 elements implies unacceptable computational times.
- sklearn pack is less automatic than TSAM, but it gives the possibility of saving more information. However, it does not allow using k-medoids, but k-means, which reminds the problem already explained.

Therefore, the following steps explain the adopted data series reduction technique, while figure 4.1.1 represents it schematically:

1. sklearn pack is used to generate clusters with k-means algorithm. The process is repeated varying the number of days, from 2 to 45. Centroids are the representative elements for the clusters. In order to keep seasons as separated as possible, the air temperature is clustered with global solar irradiance. Indeed, similar profiles for global solar irradiance, belonging to different seasons, could imply large differences in air temperature and, thus, huge differences in thermal demand. A deeper discussion about this is presented in Section 5.1.
2. For each cluster, the closest element (in terms of Euclidian distance) to the centroid is taken as new representative day. Therefore, each cluster is not represented anymore by a centroid, but by an element belonging to the cluster itself.
3. Each representative element is associated with its date, which allows coupling them with the respective profile of electricity price, electricity demand and thermal demand.
4. If a cluster includes the element with extreme conditions of thermal demand, that element becomes the representative one of the cluster. This passage is necessary, otherwise solutions are infeasible because of underestimation of thermal requirements. Clearly, the adopted solution overestimates the problem, which means that, especially for a low number of clusters, the solution is strongly conservative, as shown in Section 5.2. Indeed, extreme scenarios are not representative so, if N=2, instead of weighting one, this day could weight more than one hundred along the optimization period.
5. The number of elements belonging to each cluster gives the weight of each representative day.

The second crucial aspect is the mathematical formulation. The reference model is a Mixed Integer Linear Programming (MILP) problem. In particular, a general MILP can be formulated as:

$$\min f = \min_{x,y} c^T x + d^T y$$
$$s.t\ Ax + By \geq b$$
$$(x, y) \in \mathbb{R}^n{}_+ \times \mathbb{Z}^p{}_+ \qquad (4.1)$$

With

$$\mathbb{R}^n{}_+ = \{x \in \mathbb{R}^n : x \geq 0\}$$
$$\mathbb{Z}^p{}_+ = \{y \in \mathbb{Z}^p : y \geq 0\} \qquad (4.2)$$

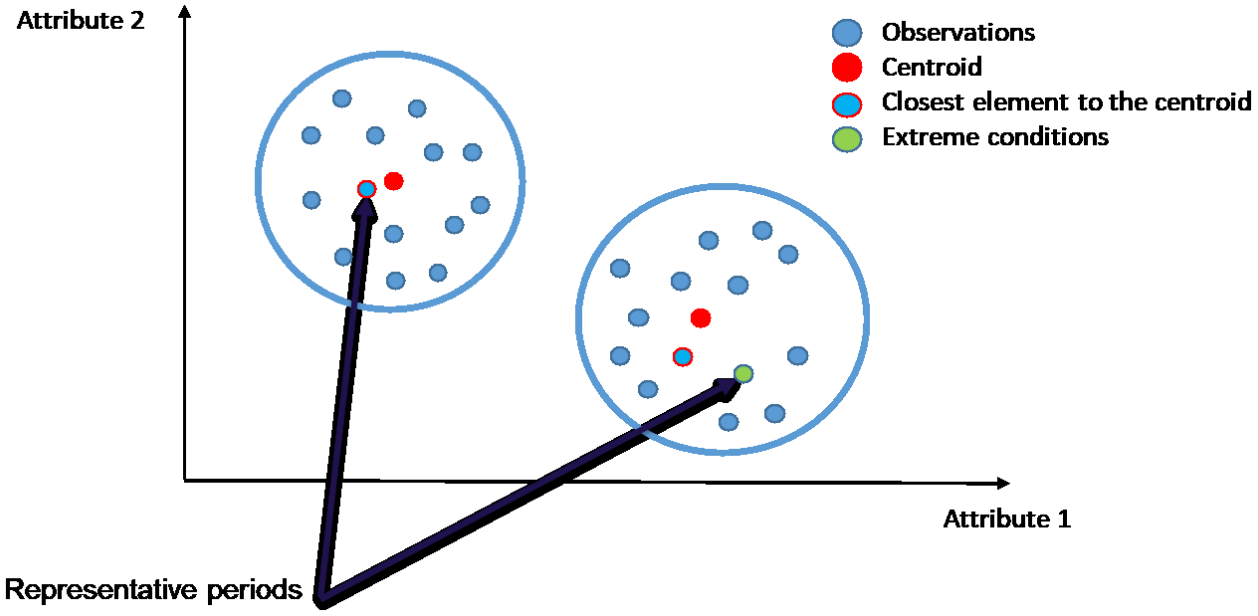

Figure 4.1.1: Simplified scheme of the clustering technique. k-means is adopted, so the centroid (red) should be the representative element. Note that it does not belong to the cluster, so the closest element to it is taken as representative element (blue and red). This is valid unless the extreme scenario (green) belongs to a cluster: if so, it becomes the representative element.

$\mathbb{R}^n{}_+$ is the n-dimensional space with all non-negative real numbers. $\mathbb{Z}^p{}_+$ is the p-dimensional space of non-negative integer numbers. Clearly, some variables in a MILP can be binary variables, thus assuming values equal to 0 or 1.

About expressions (4.1), f is the objective function, c and d are cost vectors associated with the design and operation decision variables, A and B are constraints matrixes, b is a vector with known terms.

According to Rech et al. [68], variables in a model can be independent and dependent. In particular, independent variables are set as parameters by external condition, such as global solar irradiance or the energy demand, while dependent variables are calculated as function of the independent ones.

In addition, the problem is an optimization one, which means some variables are free to vary to minimize a given objective function: these are the decision variables. In this thesis, decision variables can be divided in design and operation ones. Design decision variables are the size of the different energy conversion systems (the number of them is decided a priori) and binary variables, related to the inclusion of a given component. Differently, operation decision variables are energy fluxes for dispatchable components, such as the CHP system, and binary variables, indicating the ON/OFF status and additional variables, required to avoid bi-linear constraints.

The objective function is the minimization of the total cost. For the period 2010-2020, it can be written as

$$min\, f = min\left(Cost_{des} + \sum_{y=y_0}^{Y} \sum_{d=0}^{D} \sum_{h=0}^{H} Cost_{op,t} \right) \tag{4.3}$$

$$Cost_{des} = \frac{1}{\alpha}\big[c_{PV} * A_{PV} + \big(c_{EES,var} * E_{EES,max} + c_{EES,fix} * \delta_{EES,inv}\big) + \big(c_{CHP,var} * C_{CHP} + c_{CHP,fix} * \delta_{CHP,inv}\big) + \big(c_{boiler,var} * C_{boiler} + c_{boiler,fix} * \delta_{boiler,inv}\big) + \big(c_{TES,var} * Q_{TES,max} + c_{TES,fix} * \delta_{TES,inv}\big)\big] \quad (4.4)$$

$$Cost_{op,t} = F_{CHP,t} * c_{naturalgas} + F_{boiler,t} * c_{naturalgas} + c_{grid,purchase,t} * P_{grid,purchase,t} - c_{grid,sell} * P_{grid,sell,t} \quad (4.5)$$

Index h indicates the hour of the day, thus it goes from 0 to 23, according to the hourly resolution. Index d indicates the day and it goes from 0 to 364, covering the entire year. Index y indicates the considered year: it goes from the initial one y_0 to the final one Y. Therefore, the reference model for the testing dataset has y_0=2010 and Y=2020. If a variable varies with time, the adopted index is t, indicating the total time, including the respective hour, day and year. For example, the fuel required by the boiler is indicated as $F_{boiler,t}$ instead of $F_{boiler,h,d,y}$.

Equation (4.3) shows that the goal is to minimize the sum between design and operating costs. Equation (4.4) shows how the design cost is expressed:

1. $c_{PV} * A_{PV}$ is the product between the specific cost per area c_{PV} [€/m^2] and the area of the photovoltaic system, which is a design decision variable. c_{PV} and all the following costs are taken from an internal database, as explained in Section 5.
2. $\big(c_{EES,var} * E_{EES,max} + c_{EES,fix} * \delta_{EES,inv}\big)$ is the term related to the electric energy storage. In particular, there is a variable cost $c_{EES,var}$, multiplied for the size of the storage $E_{EES,max}$, which is a design decision variable, and a fixed term $c_{EES,fix}$, related to the inclusion of the component. In fact, $\delta_{EES,inv}$ is a binary decision variable, whose value is 1 if the component is present in the system and 0 if it is not.
3. $\big(c_{CHP,var} * C_{CHP} + c_{CHP,fix} * \delta_{CHP,inv}\big)$ expresses the design cost for the CHP system. Terms are analogue to the ones of the EES. Design decision variables are C_{CHP} and $\delta_{CHP,inv}$.
4. $\big(c_{boiler,var} * C_{boiler} + c_{boiler,fix} * \delta_{boiler,inv}\big)$ is the term for the boiler. Meanings are equal to the previous terms, so decision variables are the size C_{boiler} and the binary variable $\delta_{boiler,inv}$.
5. $\big(c_{TES,var} * Q_{TES,max} + c_{TES,fix} * \delta_{TES,inv}\big)$ represents the cost for the thermal energy storage. Design decision variables are the size $Q_{TES,max}$ and the investment binary variable $\delta_{TES,inv}$.

1/α is the term required to take into account amortization. If the optimization period is 1 year.

$$\frac{1}{\alpha} = \frac{r}{1-(1+r)^{-lifetime}} \quad (4.6)$$

The system lifetime is assumed to be equal to 20 years. r is the interest rate, equal to 4%. However, if the model is the reference one for the testing dataset, due to the consideration of eleven years instead of 20, only a part of it should be taken into account.

For what concerns the operating costs (4.6):

1. $\mathrm{F_{CHP,t}} * \mathrm{c_{naturalgas}}$ indicates the operation cost for the CHP system, equal to the fuel required for the operation $\mathrm{F_{CHP,t}}$ multiplied for the price of natural gas $\mathrm{c_{naturalgas}}$. Note that $\mathrm{F_{CHP,t}}$ is a dependent variable, because the power and heat produced, $P_{CHP,t}$ and $Q_{CHP,t}$ are the operation decision variables.
2. $\mathrm{F_{boiler,t}} * \mathrm{c_{naturalgas}}$ is the term related to the boiler. A for the CHP system, $F_{boiler,t}$ is the dependent variable.
3. $\mathrm{c_{grid,purchase,t}} * \mathrm{P_{grid,purchase,t}}$ is the cost for purchasing energy from the grid. The cost is set as explained in Section 3, while $\mathrm{P_{grid,purchase,t}}$ is a second-stage decision variable.
4. $\mathrm{c_{grid,sell}} * \mathrm{P_{grid,sell,t}}$ is the earning obtained by selling energy to the grid. In fact, it has a negative sign in the objective function. $\mathrm{P_{grid,sell,t}}$is a decision variable.

The previous objective function, unfortunately, is valid with 93690 time steps: if N representative days are generated and used as optimization period, the objective function has to be changed. Indeed, it becomes

$$min\, f = min\left(Cost_{des} + \sum_{d=0}^{D} w_d * \sum_{h=0}^{H} Cost_{op,t}\right) \tag{4.7}$$

With w_d weight associated to the specific representative day. As explained before, it is equal to the number of elements belonging to a given cluster. Now d goes from 0 to D=N, the number of representative days obtained through clustering techniques. The index for the variable remains t, the total time.

In addition, constraints are required to solve the problem. Energy balances are presented for both electrical and thermal side. Additionally, for each energy conversion unit, characteristic curves are included, as well as inequalities for their operability range. Mass balances are not indicated, to keep the number of equations as low as possible, because they are assumed to be verified once energy balances are. All these constraints are presented in Section 4.2.

This model, as already said, is used as reference. The following procedure explains in detail all the passages:

1. N representative days, with $N \in [2, 45]$, are obtained through clustering techniques. Consequently, 44 set of representative days are generated.
2. For each set of days, the model is solved, obtaining the optimized solution. In particular, one set of sizes, the design decision variables, is considered for each set of representative days.
3. Each set of sizes is tested for the period 2010-2020, as described in Section 4.5 (testing dataset), which means the design decision variables are set, thus only operation is optimized. This is a test on real data, with 93690 time steps. It is possible because of the lower computational effort, linked with the absence of design decision variables.
4. Figure 5.2.1 shows the graph obtained. For each set of sizes in the x-axis, associated with a solution obtained with $N \in [2,45]$ representative days, a point on the y-axis, indicating the total cost for period 2010-2020, is coupled. It represents the real total cost for the entire period if the sizes in the design phase are those specific ones.
5. The minimum for the graph in figure 5.2.1 is chosen as the reference solution. Indeed, the objective function is the minimization of the total cost and, because the solutions found in point 2 are tested with the real data, there are no problems about loss of information due to data series reduction. Clearly, this solution is not the reference one, but a higher bound.

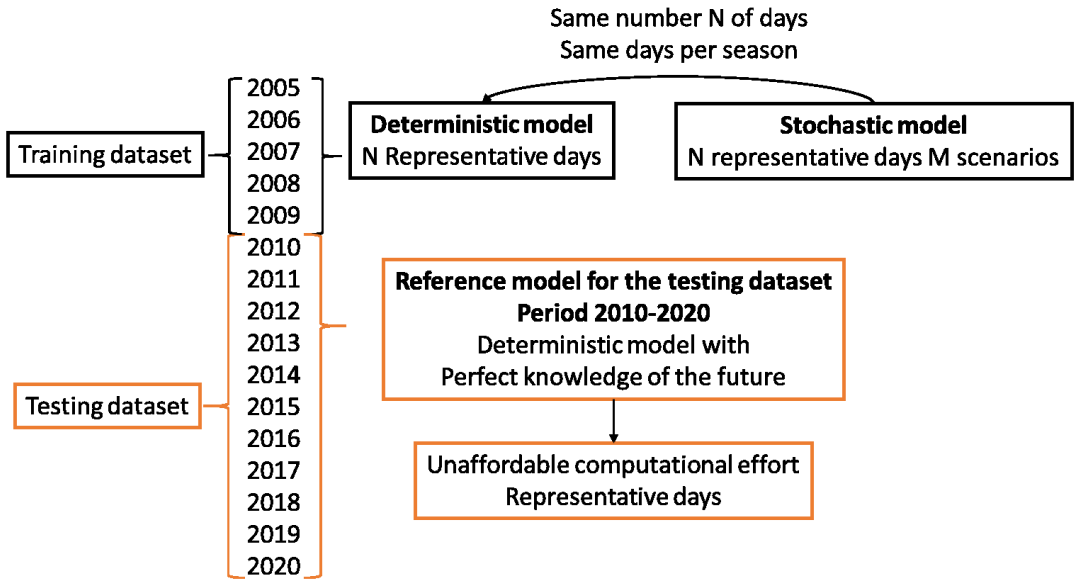

Figure 4.1.2. Summary of the different models for the residential MES. The period 2005-2009 represents the training dataset, while the period 2010-2020 the testing one. For the training dataset, a deterministic model with N representative days and a two-stage stochastic model with N days and M scenarios are developed. They are compared on this period and on the testing dataset, with respect to the reference model.

4.2 Energy conversion systems

4.2.1 Photovoltaic system

The electric power from the photovoltaic system is given by equation (4.8), according to [69]:

$$P_{PV,t} = A_{PV} * G_{tot,t} * \eta_{PV,t} * \eta_{BOS} \quad (4.8)$$

Here, A_{PV} is the area of the PV system, a design decision variable. $G_{tot,t}$ is the global solar irradiance on the tilted plane, assumed to have optimal tilt and azimuth angles: it is the uncertain parameter. $\eta_{PV,t}$ is the efficiency of the PV system, which depends on the air temperature, as shown in equations (4.9, 4.10). η_{BOS} is the balance of system efficiency, whose value is set.

$$T_{cell,t} = T_{air,t} + \frac{NOCT - 20}{800} * G_{tot,t} \quad (4.9)$$

$$\eta_{PV,t} = \eta_{PV,std} * \left(1 - b_0 * \left(T_{cell,t} - 25\right)\right) \quad (4.10)$$

4.2.2 Electric energy storage system

$$E_{EES,t} = E_{EES,t-1} * (1 - E_{selfdischarge}) + \eta_{EES} * P^{+}_{EES,t} - \frac{1}{\eta_{EES}} P^{-}_{EES,t} \quad (4.11)$$

This equation shows how the energy in the storage at time t is linked to the energy at the previous moment. In particular, it is reduced by a term equal to $(1 - \mathrm{E_{selfdischarge}})$, considering the self-discharge capacity of the system, reduced by the energy taken from it and increased of the energy charged. Clearly, η_{EES} indicates the efficiency of the storage. If a quantity of energy $\mathrm{P^{-}_{EES,t}}$ is required, $\frac{1}{\eta_{\mathrm{EES}}} \mathrm{P^{-}_{EES,t}}$ is discharged from the battery, to consider losses. Similarly, if a quantity $\mathrm{P^{+}_{EES,t}}$ could be charged to the storage, only a smaller amount $\eta_{\mathrm{EES}} * \mathrm{P^{+}_{EES,t}}$ is charged in the end.

$$SOC_{min} E_{EES,max} \leq E_{EES,t} \leq SOC_{max} E_{EES,max} \quad (4.12)$$

$$E_{EES,0} = 0.5 * E_{EES,max} \quad (4.13)$$

$$E_{EES,T} = 0.5 * E_{EES,max} \quad (4.14)$$

$$0 \leq P^{+}_{EES,t}, P^{-}_{EES,t} \leq const * E_{max,EES} \quad (4.15)$$

Equation (4.12) forces the energy level to be included between the minimum and maximum state of charge SOC. (4.13) and (4.14) are boundary layers, indicating the initial and final amount of energy in the storage. Equation (4.15) indicates that charge and discharge power cannot be higher than a predefined value, which is function of the size of the storage.

In addition, two equations are required to force the storage to charge OR discharge in a given moment, preventing the two processes simultaneously. The big M method is used for this:

$$P_{EES,discharge,t} \leq M_{BIG} * \delta_{EES,t} \quad (4.16)$$

$$P_{EES,charge,t} \leq (1 - M_{BIG}) * \delta_{EES,t} \quad (4.17)$$

Here $\mathrm{M_{BIG}}$ is a parameter, whose value is set and very big, in order not to reach it. δ_{EES} is a binary decision variable: if it is equal to one, the EES is in discharge mode, if it is 0 the EES is charging energy. Another equation is required, to force $\delta_{EES,t}$ to be equal to zero at any time if the EES is not included.

$$\delta_{EES,t} \leq \delta_{EES,inv} \quad (4.18)$$

4.2.3 CHP system

CHP characteristic curves are obtained through interpolation from data given by [61]

$$F_{CHP,t} = 0{,}2722 * P_{CHP,t} + 3{,}7125 * \delta_{CHP,t} \quad (4.19)$$

$$Q_{CHP,t} = 1{,}7246 * P_{CHP,t} + 9{,}3109 * \delta_{CHP,t} \quad (4.20)$$

Equation (4.19) links the fuel consumption at time t $F_{CHP,t}$ to the power generated in the same moment $P_{CHP,t}$, which is a decision variable. Clearly, a binary variable $\delta_{CHP,t}$ is required to connect them properly: if the system is working, $\delta_{CHP,t}$ is equal to one, otherwise the system is off and $\delta_{CHP,t}$ is zero. Similarly, equation (4.20) links the heat produced by the system to the power produced.

The system is free to choose the size of the CHP plant. Therefore, to avoid bilinear constraints [70] as

$$P_{CHP,t} * \delta_{CHP,t} \leq C_{CHP} \quad (4.21)$$

In which there is a product between two decision variables, increasing exponentially the complexity of the code, an additional variable $\vartheta_{CHP,t}$ is added, as shown in the following equations

$$\delta_{CHP,t} * C_{CHP,min} \leq \vartheta_{CHP,t} \quad (4.22)$$

$$\vartheta_{CHP,t} \leq \delta_{CHP,t} * C_{CHPmax} \quad (4.23)$$

$$\left(1 - \delta_{CHP,t}\right) * C_{CHP,min} \leq C_{CHP} - \vartheta_{CHP,t} \quad (4.24)$$

$$SOP_{CHP,min} * \vartheta_{CHP,t} \leq P_{CHP,t} \quad (4.25)$$

$$\vartheta_{CHP,t} \leq P_{CHP,t} \quad (4.26)$$

Note that, in any case, even if the system can decide the optimal size, it is not possible to accept every size. In fact, it depends on the catalogue [61], which means the closest one to a possible model must be considered.

The last equation (4.27), as for the EES, links the binary variable $\delta_{CHP,t}$ to the investment one $\delta_{CHP,inv}$, required in the objective function (Section 4.1). In this way, if the component is not included, there is no possibility of producing power or heat.

$$\delta_{CHP,t} \leq \delta_{CHP,inv} \quad (4.27)$$

4.2.4 Boiler

$$F_{boiler,t} = Q_{boiler,t}/\eta_{boiler} \quad (4.28)$$

Equation (4.28) links the fuel consumption $F_{boiler,t}$ to the heat produced $Q_{boiler,t}$ (decision variable) through the efficiency of the component η_{boiler}. As for the CHP system, the size is not set, entailing the introduction of the additional variable $\vartheta_{boiler,t}$, as indicated in equations (4.29-4.34).

$$\delta_{boiler,t} * C_{boiler,min} \leq \vartheta_{boiler,t} \quad (4.29)$$

$$\vartheta_{boiler,t} \leq \delta_{boiler,t} * C_{boiler,max} \quad (4.30)$$

$$\left(1 - \delta_{boiler,t}\right) * C_{boiler,min} \leq C_{boiler} - \vartheta_{boiler,t} \quad (4.31)$$

$$SOP_{boiler,min} * \vartheta_{boiler,t} \leq P_{boiler,t} \quad (4.32)$$

$$\vartheta_{boiler,t} \leq P_{boiler,t} \quad (4.33)$$

$$\delta_{boiler,t} \leq \delta_{boiler,inv} \quad (4.34)$$

4.2.5 Thermal energy storage system

$$Q_{TES,t} = Q_{TES,t-1} * (1 - E_{selfdischarge}) + \eta_{TES} * Q^{+}_{TES,t} - \frac{1}{\eta_{TES}} Q^{-}_{TES,t} \quad (4.35)$$

$$SOC_{min} Q_{TES,max} \leq Q_{TES,t} \leq SOC_{max} Q_{TES,max} \quad (4.36)$$

$$Q_{TES,0} = 0.5 * Q_{TES,max} \quad (4.37)$$

$$Q_{TES,T} = 0.5 * Q_{TES,max} \quad (4.38)$$

$$0 \leq Q^{+}_{TES,t}, Q^{-}_{TES,t} \leq const * Q_{max,TES} \quad (4.39)$$

$$Q_{TES,discharge,t} \leq M_{BIG} * \delta_{TES,t} \quad (4.40)$$

$$Q_{TES,charge,t} \leq (1 - M_{BIG}) * \delta_{TES,t} \quad (4.41)$$

$$\delta_{TES,t} \leq \delta_{TES,inv} \quad (4.42)$$

Equations are analogue to the ones of the EES. Indeed, (4.35) links the energy level in the storage at time t to the one at t-1, (4.36) forces the system to stay in the range between minimum and maximum state of charge level, (4.37) and (4.38) state the initial and final energy level for the storage. Equation (4.39) indicates that the power charged/discharged is a function of the size of the TES. Big M method is applied in equations (4.40), (4.41) to avoid a charge/discharge process in the same time step. As for the other components, inequality (4.42) is required to connect the binary variable $\delta_{TES,t}$, used for ON/OFF status, to the design decision binary variable $\delta_{TES,inv}$.

4.2.6 Grid

$$P_{grid,sell,t} \leq \delta_{grid,t} * M_{BIG} \quad (4.43)$$

$$P_{grid,purchase,t} \leq \left(1 - \delta_{grid,t}\right) * M_{BIG} \quad (4.44)$$

Expressions (4.43), (4.44) are provided to prevent the grid to sell to the system and buy from it in the same moment, by using the big M method.

4.2.7 Energy balances

The electric energy balance is, for each moment

$$P_{PV,t} + P_{CHP,t} + P_{grid,purchase,t} + P^{-}_{EES,t} = P^{+}_{EES,t} + P_{grid,sell,t} + P_{load,t} \quad (4.45)$$

Energy that flows into the system comes from the PV plant or the CHP one, is purchased from the grid or is taken from the EES. On the contrary, energy that leaves the system is charged in the EES, sold to the grid or required by users.

The thermal balance states

$$Q_{CHP,t} + Q_{boiler,t} + Q^{-}_{TES,t} = Q^{+}_{TES,t} + Q_{load,t} \quad (4.46)$$

Hence, the CHP plant, the boiler and the heat stored in the thermal storage are able to satisfy the thermal load. On the contrary, if heat production is too much, it can be stored in the TES.

Please note that previous equations are referred to the reference model for period 2010-2020. If n representative days are used, instead, most of equations are the same. In any case, equations (4.14), (4.28) change a little, because these N days are not consecutive and aggregate periods for different months and years. Therefore, all of them are independent from each other, forcing the electric and thermal energy storages to have the same energy level at the beginning and at the end of every day.

4.3 Deterministic model of the residential MES with N representative days, for period the 2005-2009

4.3.1 Decision variables, constraints and objective function

This model, as the one presented in Section 4.1, is referred to the residential MES introduced in Section 2: the goal is to see how close solutions can be with respect to the ones from the reference model with perfect knowledge about the future. In particular, it considers as input data global solar irradiance, air temperature, price of energy carriers, electrical energy demand, thermal demand and techno-economic characteristics for the energy conversion systems. Moreover, constraints for the model, implying characteristic curves for each energy conversion unit and inequalities, as well as output variables, are the same. The objective function is the one presented at the end of Section 4.1 (equation (4.7)), because of the N days as optimization period. Hence, the difference is given by how global solar irradiance and air temperature are obtained.

In particular, differently from the previous model, even if the global solar irradiance is taken from the same source, the period is different (2005-2009): the final goal is to see how close solutions can be with respect to the best ones, in terms of total cost. Therefore, once generated the clusters, an optimization process is done, in order to get design and operation decision variables. Then, sizes for energy conversion units are set and operation can be tested both on the training dataset, thus period 2005-2009 (43800 time steps), or testing dataset, period 2010-2020 (93690 time steps).

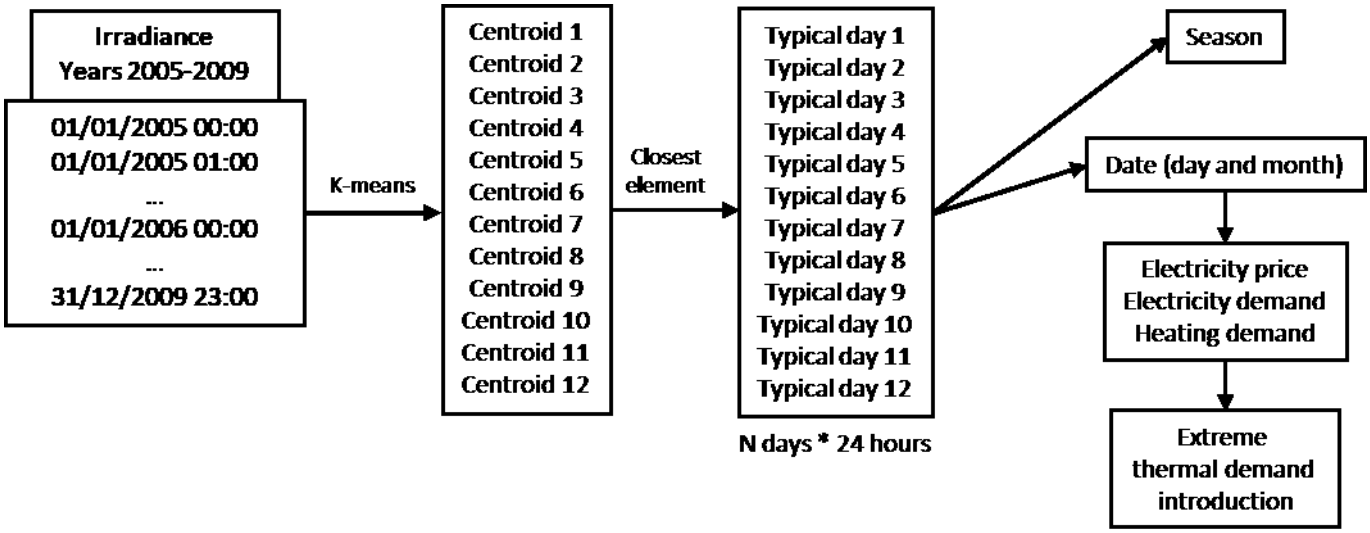

Figure 4.3.1: Input data for the deterministic model with N representative days, generated with clustering techniques. In this example, N is equal to twelve.

4.3.2 Generation of N representative days with clustering techniques

Clustering techniques are adopted to reduce computational effort and obtain days (clusters) as representative as possible of the initial dataset.

The procedure is similar to the one already presented in the previous Section:

1. Starting from the global solar irradiance and air temperature for the period 2005-2009, with hourly resolution, the clustering process, identical to the one explained in the previous Section, gives 29 sets of N representative days, with $N \in [2, 30]$.
2. The optimized solution, implying design and operation decision variables, is found for each set of clusters. The optimization period corresponds to 24*N time steps.
3. If the total cost, obtained in point 2, is plotted as function of the number of representative days, a curve is obtained. However, this cost depends on the adopted clustering technique. Indeed, as already indicated, the introduction of extreme scenarios force to overestimate the total cost.
4. Therefore, for each set of sizes the operation is optimized for period 2005-2009. The minimum of the curve with actual total cost on the y-axis and set of sizes, linked to the representative days, on the x-axis, gives the optimal size for the components and the optimal total cost.
5. The simulation is repeated many times, because results vary depending on the clustering process. The chosen number of representative days should be as low as possible, to limit computational time, and as close as possible to the minimum of graph described in point 4.

The alternative is to choose the number of representative days with respect to the yearly reference models, always describing the same residential MES. Input data is the same, but values for the air temperature and the global solar irradiance are related to the specific year (period 2005-2009). Equations, inequalities and the output variables are the same, but the number of time steps, equal to 8760, which means computational effort is still acceptable, is the difference. Therefore, the model gives the best solution for the specific year, which is surely the reference.

If the comparison is done in this way, two alternatives are possible. The first one is to do as just explained, considering the entire training dataset in a single clustering process to get a limited number of representative days.

The other possibility is to compare solutions year by year. The procedure can be summarized as:

1. The reference model for each year is solved. Hence, the best solution in terms of design decision variables and objective function is found.
2. For every year of the training dataset, the clustering process used up to now is used, to generate 29 sets of days for each year, for a total of 29*5=145 sets. Therefore, five sets are made up of 2 representative days, five sets of 3 days, etc. Considering sets with 2 days, the first one is referred to 2005, the second to 2006 and so on.
3. These models with a limited number of days are optimized and their solutions are compared to the one of the reference model for the respective year.

In order to compare these results to the reference one, two types of indicators are taken into account. The first one is the Mean Absolute Error (MAE), simply calculated as

$$MAE_N = \frac{\sum_{i=2005}^{2009} \sum \left| y_i - x_{i,N} \right|}{Y} \qquad (4.47)$$

Where $x_{i,N}$ is the value of the objective function (total cost) for the different models with N representative days, related to year i, while y_i is the objective function for the reference model of year i. Y is the number of years, equal to five for the training dataset.

The second indicator is the Root Mean Square Difference (RMSD), evaluated as

$$RMSD_N = \sqrt{\frac{\sum_{i=2005}^{2009}\left(y_i - x_{i,N}\right)^2}{Y}} \qquad (4.48)$$

The meaning of the different elements of (4.48) is the same as for (4.47).

Once obtained these values for each year i, the number of days is chosen by considering the minimum values of MAE and RMSD. Then, the actual days are the ones of year i, with the lowest values of MAE and RMSD. For example, if MAE and RMSD are minimized with k=20, these 20 days are the ones from the closest year with respect to the reference model for the same year.

Please note that the model is the same in terms of mathematical formulation and output variables: the only difference is given by how representative days are generated. In particular, this procedure allows obtaining data for global solar irradiance and air temperature, with respect to one year, while other input, such as electricity price and energy demands, are related to the training dataset or the testing one. As for the previous methods, date is obtained for each representative day and this allows coupling the inputs.

4.3.3 Alternative methods to generate representative days based on seasons

In Section 4.3.2, clustering process considers both global solar irradiance and air temperature. The idea is to keep seasons as separated as possible, without rigid distinctions.

The motivation for separating seasons for clusters is not related to a succession of data. In fact, once obtained the set of representative days, there is not an historical progression of elements, which means that all scenarios are independent from each other. This forces the system to find a worse solution because, as indicated in Sections 4.2.2, 4.2.5, energy storage systems relate their energy level to the one of the previous time step, the previous hour. If each scenario is independent, energy level at hour 00:00 cannot be related to the one at hour 23:00, but must be set.

Hence, the consideration of just climate conditions as uncertain parameters is the key: assuming to work just with global solar irradiance, if a cluster is represented by a summer element, for instance 18/07, but half of its elements are from winter, it will not entail a good solution. Indeed, the global solar irradiance for this cluster represents well the elements, but the remaining input data (energy demand, air temperature, electricity price) is related just to the representative day, thus 18/07. Therefore, global solar irradiance could be similar for summer and winter days, depending on several aspects, but air temperature, for instance, will be probably different.

Global solar irradiance does not change sharply once the season changes, but it depends on several factors. Hence, some clusters near the seasonal “border” could be well represented by half elements from a season and half from the previous/following one. Because of this, the choice was not to separate into seasons, in order to let the system decide how to aggregate elements. Nevertheless, to help the process, air temperature is included in the clustering process.

In order to have a complete view about the problem, seasonal representative days are obtained. In Section 5, comparison of results is deepened. Note that the only difference is given by how representative days are obtained.

1. The training dataset is divided per season. Therefore, winter section includes winter elements for 2005, 2006, 2007, 2008 and 2009.
2. For each season, a clustering process is carried out. The procedure is always the same, with a number N of representative days, with $N \in [1, 7]$. Therefore, 4*N representative days are generated, N per season. Please note that combinations of different number of days per season are not considered to avoid computational complexity. Indeed, the quality of clusters is not a proper indicator for the quality of the solution (Section 4.4), so the decision could be done just testing all the sets of days to obtain optimized solutions.

Another method to generate representative days is to obtain average monthly profiles. In particular, the training dataset (2005-2009) is divided per month and, given the hour, the day and the month, the average values of these parameters are calculated. For example, if irradiance and temperature of January 2005, 2006, …, 2009 are taken into account

$$G_{h,d,January} = \sum_{y=2005}^{2009} G_{h,d,y,January}/5 \qquad (4.49)$$

With $h \in [0, 23]$ hour of the day, $d \in [0, D]$ day of the specific month, $y \in [2005, 2009]$ year.

The clear disadvantage in using this method is the averaging of curves. Indeed, an average profile for global solar irradiance and air temperature implies a flatter curve, which avoids extreme hot and cold conditions and entails overestimation or underestimation of design variables. In fact, an alternative would be to consider average seasonal days, dividing the dataset per season instead of month. However, this solution would imply even flatter curves, with a number of representative days equal to four. In addition, the rigid distinction linked to seasons is still a problem.

4.4 Two-stage stochastic model of the residential MES with N representative days and M scenarios, for the period 2005-2009

4.4.1 Decision variables, constraints and objective function

The two-stage stochastic model is always related to the same scheme presented in Section 2. The aim is to compare its solution, in terms of total cost and design decision variables, with the ones from the reference model for the period 2010-2020 (Section 4.1) and from the deterministic one with N representative days (Section 4.3).

The number of representative days is the same as the one for the deterministic model. The difference, however, is the generation of M scenarios of global solar irradiance and air temperature for each representative day.

As explained in Section 1.6, the idea is to make a decision, whose convenience depends on an outcome at time 1, at time 0. The here and now approach is adopted, minimizing the expected value of the objective function, which has elements of both first and second stage.

$$z_{HN} = \min_{x} \mathbb{E} f(x, s) \tag{4.50}$$

Indeed, the approach is to consider the first stage for the design, which is done at time 0. Then, each representative day (time 1) will have a set of scenarios, weighted with its own probability. Therefore, the operation is represented by one single stage, even if it occurs for one or more years. The choice is not to add other stages unless information about future changes, which is not the case because of the consideration of climate conditions as uncertain parameters, avoiding climate change introduction [9]. Because of that, the objective function becomes:

$$min\, f = min\left(Cost_{des} + \sum_{s=0}^{S} p_s * \sum_{d=0}^{D} w_d * \sum_{h=0}^{H} Cost_{op,t} \right) \tag{4.51}$$

In which s indicates the specific scenario, S=M the number of scenarios, p_s the probability associated with each scenario. As can be noted, the design cost does not depend on the scenario s, but the convenience in installing an energy conversion unit depends on that. Therefore, the design decision variables depend on the expected value for the operation.

Constraints are equal to the one already presented for the deterministic model (Section 4.2), so with respect to equations (4.8-4.46), except for equations (4.14), (4.28) because each day (and each scenario for the day) have to be independent from each other. In addition, decision variables are the same as indicated in Section 4.1. In addition, input data presented in Section 3 does not vary.

Therefore, the focus should be on irradiance and temperature. Following sub-Sections describe:

1. Evaluation of representative days.
2. Evaluation of scenarios for each day.

4.4.2 Proposed method to generate representative days and stochastic scenarios

As explained in Section 4.3.2, the number of days for the deterministic model, considering computational burdens and error with the best solution. According to results in Section 5.3, it can be set as N=9 or 11. Let us suppose to work with 11 day, to get a number of days closer to the ones obtained with other methods and to have a fair comparison. A real day represents each cluster: its belonging season is found. Indeed, to compare solutions it is required to have the same days for each season, for both the deterministic solution and the two-stage stochastic one. If it is not checked, an explicative case can be taken as example: from the deterministic model, 6 days are from summer, while in the stochastic one 6 days are from winter. Even if profiles are just for solar irradiance and air temperature, representative elements of each cluster are coupled with profiles of electricity price, electricity demand and thermal demand. Therefore, under these assumptions the deterministic model will overestimate the PV production, while underestimating CHP, TES and boiler sizes, while the stochastic one will do the opposite.

This aspect is not trivial, because every time the clustering process is done clustering results are different, implying changes in representative days and in days per season. In fact, table 4.4.1 shows the representative days for 10 simulations: 25th of February, for instance, is used as representative day just four times out of ten simulations. Table 4.4.2 indicates that number of representative days per month vary for each simulation. In the same way, the number of elements per season may vary, even if it tends to be more constant.

Because of this, the dataset is divided for the different seasons. Then, a first clustering process is done, as explained in Section 4, using k-means method with sklearn pack. To be precise, four clustering processes are done, one per season: for example, focusing on winter, if the deterministic model has three representative days in that period, k-means is carried out with number of clusters equal to three. Then, the centroid is substituted with its closest element. The same is done for the other seasons, obtaining N=11 representative days.

Representative days deterministic model																			
Simulation 1		Simulation 2		Simulation 3		Simulation 4		Simulation 5		Simulation 6		Simulation 7		Simulation 8		Simulation 9		Simulation 10	
25	2	18	7	23	1	28	8	18	7	20	5	14	2	24	10	18	7	10	3
28	9	28	8	6	5	29	8	14	1	20	11	28	9	29	8	20	11	23	1
20	11	10	3	10	3	15	2	6	5	24	10	19	11	20	11	9	10	20	8
20	8	23	1	29	10	19	11	28	9	22	4	10	7	25	2	23	1	14	1
24	10	20	11	28	9	10	3	28	8	29	10	26	12	18	7	6	5	25	5
10	7	29	8	28	8	30	4	23	1	10	3	23	1	10	3	25	2	10	7
23	1	6	5	30	4	23	1	20	8	23	1	15	3	16	8	20	5	29	10
20	5	14	1	20	11	18	7	20	11	20	8	20	5	9	10	29	8	22	4
16	8	25	5	18	7	20	5	29	10	16	8	29	10	20	5	12	3	28	8
10	3	19	11	19	11	9	12	24	10	28	9	22	8	23	1	29	10	20	11
29	10	28	9	29	8	6	5	10	3	10	7	22	4	29	10	24	10	28	9
22	4	29	10	20	5	28	9	20	5	25	2	20	8	6	5	16	8	19	11

Table 4.4.1: Typical days for the deterministic model repeating the clustering process ten times. Every time it is repeated, representative days vary, implying differences in the solution.

After this first step, for each typical day, another clustering process is done, in order to get its different scenarios, whose number is set arbitrarily equal to $M \in [2,7]$. Therefore, N clustering processes are done for the second step, one per representative day, plus four in the first step, one per season.

	Sim 1	Sim 2	Sim 3	Sim 4	Sim 5	Sim 6	Sim 7	Sim 8	Sim 9	Sim 10
Jan	1	2	1	1	2	1	1	1	1	2
Feb	1	0	0	1	0	1	1	1	1	0
Mar	1	1	1	1	1	1	1	1	1	1
Apr	1	0	1	1	0	1	1	0	0	1
May	1	2	2	2	2	1	1	2	2	1
Jun	0	0	0	0	0	0	0	0	0	0
Jul	1	1	1	1	1	1	1	1	1	1
Aug	2	2	2	2	2	2	2	2	2	2
Sep	1	1	1	1	1	1	1	0	0	1
Oct	2	1	1	0	2	2	1	3	3	1
Nov	1	2	2	1	1	1	1	1	1	2
Dic	0	0	0	1	0	0	1	0	0	0

Table 4.4.2: Number of representative days per month, repeating the clustering process ten times.

Please note that these processes are done just for the global solar irradiance and air temperature: once known the representative days, their date is found, in order to couple them with respective demand and electricity price curves. However, instead of the respective profiles for climate conditions, four scenarios, corresponding to different days, indicate them. On the contrary, if electricity price and energy demand curves were coupled with global solar irradiance (and temperature) of different scenarios, the uncertain parameters would not be just climate conditions anymore, because demand and electricity price would change as well.

Let us stress this point. If number of scenarios M is equal to 4 and number of representative days N is equal to 11, the comparison cannot be done between the deterministic with 11*4=44 representative days and the stochastic one with 11 days and 4 scenarios. In fact, the deterministic with 44 days has 44 profiles of global solar irradiance, of electricity price and of demand as input. Hence, these are “rigid” scenarios, differently from the stochastic model, where 44 profiles of solar irradiance are adopted, but just 11 profiles for the other parameters (assumed certain) are considered. This concept is represented in figure 4.4.3.

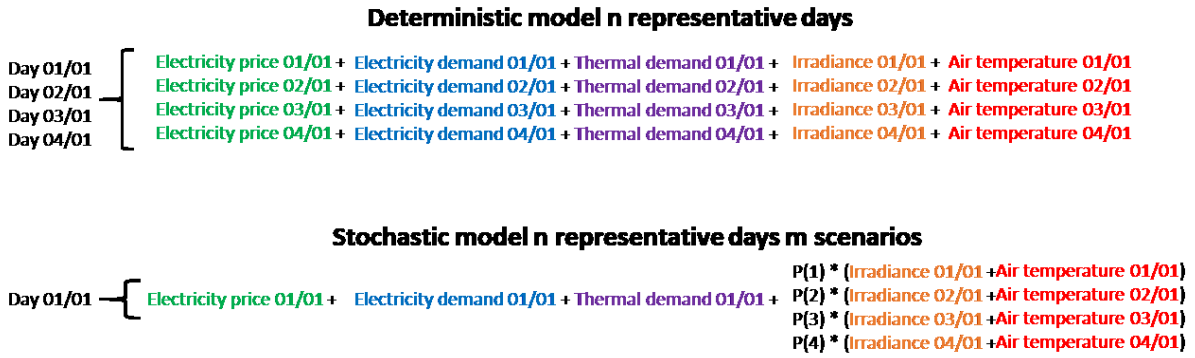

Figure 4.4.3: Association of input parameters for the deterministic and the stochastic model.

From figure 4.4.3, let us suppose one of the eleven days is the 1st of January, because it is the closest element to the centroid of the respective cluster. Now, a second clustering process, involving the same uncertain parameters, is accomplished and gives as four scenarios 1st, 2nd, 3rd and 4th of January: each of them will have its own irradiance and temperature profile, but the same electricity price (and demand curves), corresponding to the 1st of January. The two-step clustering process is indicated in figure 4.4.4.

The weight of each typical day is the number of elements in the respective cluster divided per five, because the total cost is taken back to one year, while the historical data refers to five years in a row. Then, the probability for each scenario is equal to the ratio between the number of elements belonging to that specific scenario and the total number of elements for the typical day. Note that the sum of probabilities for each day (M scenarios) has to be equal to one.

$$weigth_d = \frac{n_{elements} \in d}{5} \quad \forall d \in [0, N] \tag{4.52}$$

$$p_{d,s} = \frac{n_{elements} \in s}{n_{elements} \in d} \; \forall s \in [0, M] \; \forall d \in [0, N] \tag{4.53}$$

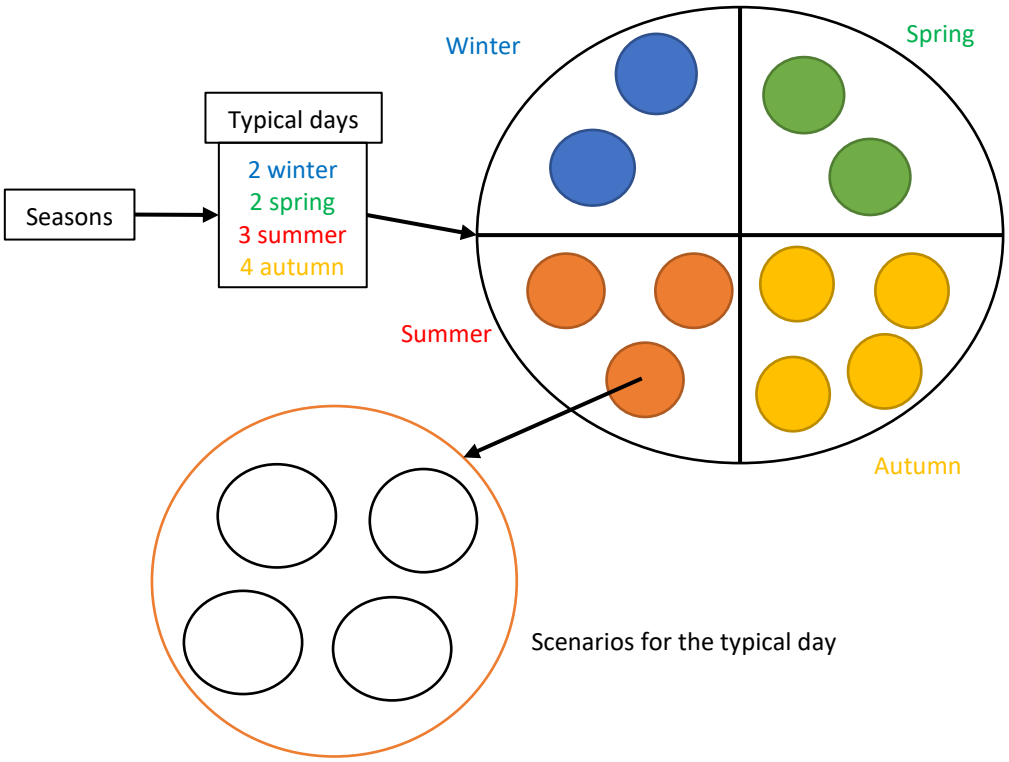

Figure 4.4.4: Generation of representative days and scenarios for the two-stage stochastic model.

4.4.3 Alternative method based on quality of clusters

Up to now, the number of clusters, required a priori by k-means algorithm, is chosen by considering the quality of the solution. Indeed, in Section 4.1 and 4.3 a variable number of representative days N is tried in order to obtain the best possible solution. However, no discussion is done about quality of clusters.

The opposite method would be to choose the number of clusters basing on the clusters quality, thus still a priori, but obtained before solving the respective model. The presented method considers four typical days instead of twelve and a variable number of scenarios: instead of starting from the deterministic method to

get the clusters numbers, the two-stage stochastic model gives the number of clusters to be used in the deterministic model.

The idea is still to divide the dataset in the four different seasons to obtain seasonal clusters with k-means method and, then, to consider all possible combinations from seasonal days. Finally, another k-means process is done in order to reduce final clusters, corresponding to typical years, made up of four typical days.

Because of the necessity of deciding the number of clusters a priori, several tentatives have to be done in order to choose high quality clusters. In order to evaluate it, for each season the process is repeated to create a profile of average silhouette value (ASV), as can be seen, for instance, for winter season.

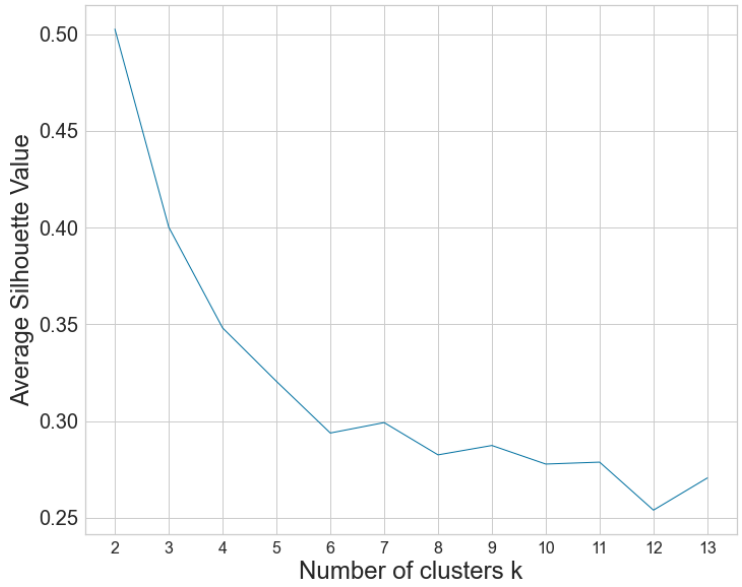

Figure 4.4.5. Average silhouette value for winter season. The value is obtained for a number of clusters from 2 to 13.

As already exaplined in Section 1.2.6, if the number of clusters k is too low, the ASV will be low, because $a_{sim}(i)$ cannot be high, due to differences of elements in the same cluster. On the contrary, if the number of clusters is too high, ASV is low, because $a_{sim}(i)$-$b_{dis}(i)$ will be low, due to the fact that an element can stay both in cluster A or B.

Returning to figure 4.4.5, it seems clear the best choice is to have k=2. Another confirmation is given by figure 4.4.6, which represents the value of sil(i) for the elements in each cluster for k=2, 6 or 12 for winter case.

As just explained, the ASV is the highest for k=2. Additionally, considering single elements, the number of them with sil(i) lower than 0 strongly increases with the increase of k. In any case, solutions with k=6 or 12 are considered, because they allows having a higher variability of scenarios

Please note that this process allows obtaining clusters for a specific season, but the operation must consider all the year, which means that combinations among seasonal typical days have to be done. In particular, if k=2, the number of combinations is equal to 16: these will simply be the final scenarios, as indicated in figure 4.4.6. On the contrary, with k=6 or k=12, the number of combinations is respectively equal to 1296 or 20736, forcing a second k-means step, in which the number of elements is arbitrary chosen, due to the inconsistency of the ASV profile (figure 5.5.7), which is always very low and does not have a “elbow” curve. In this case, number of clusters is set equal to 12, because the ASV range is very low and, in this way, a higher number of scenarios is taken into account.

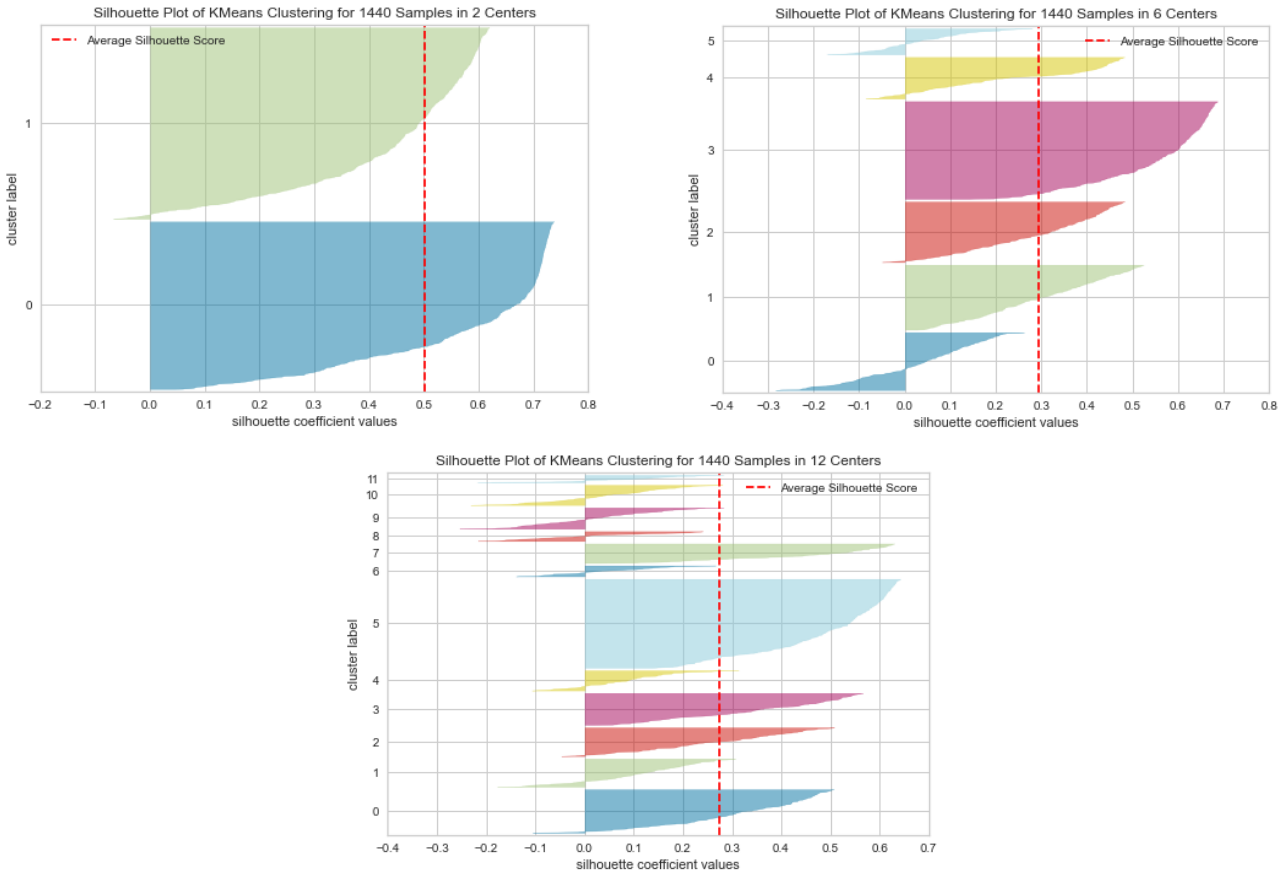

Figure 4.4.6. Silhouette plot, winter case, for number of clusters k=2, 6 or 12.

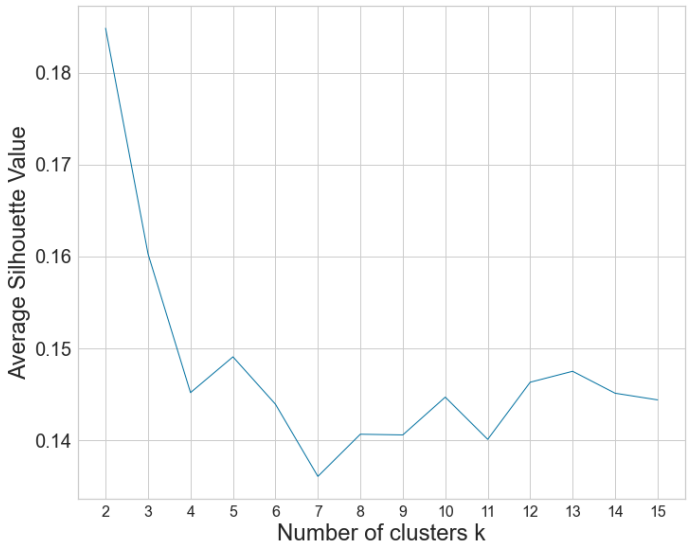

Figure 4.4.7. ASV curve for typical years. Please note that, even with k=2, the value remains under 0.2

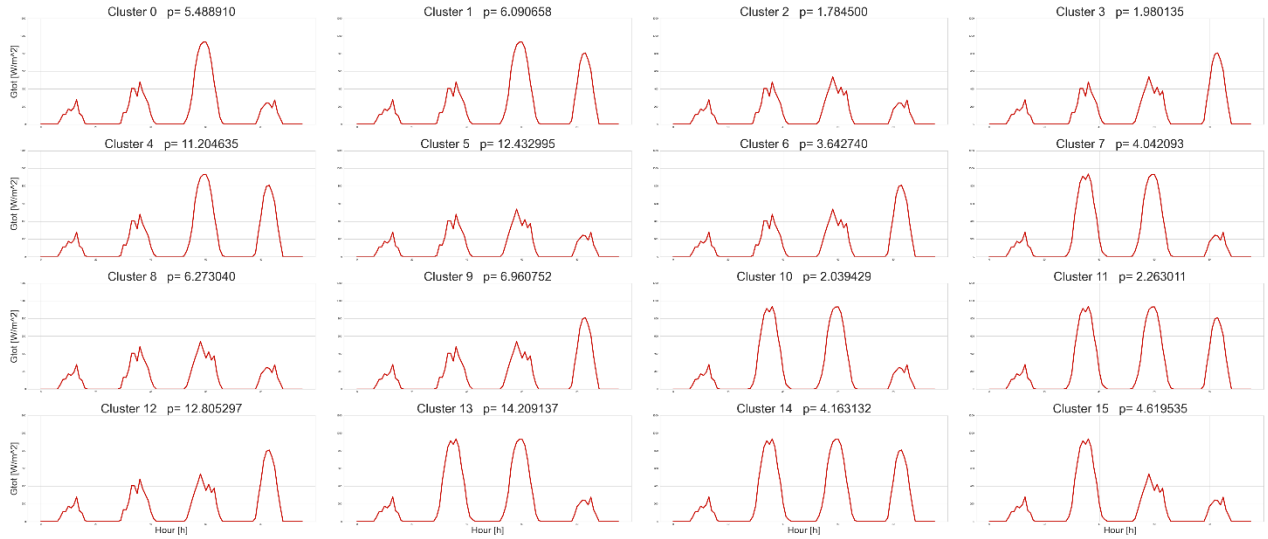

Figure 4.4.8. Scenarios, indicating a typical year, starting from seasonal k=2, with their probabilities indicated

Another crucial aspect is the determination of probabilities for each typical year. In fact, for seasonal clusters, it is sufficient to calculate cluster frequencies, defined as the ratio between the number of elements in a cluster and the total number of elements. In case of k=2, combinations are 16 and there is no need of a second clustering process, which implyes a probability for each cluster equal to the product among all seasonal probabilities (equation (4.54)).

Probability evaluation for k=6, 12 is more complex, because it has to take into account seasonal clusters probability $p_{seasonal,tot,k}$ and yearly clusters one, $p_{yearly,k}$. Hence, yearly clusters probability is calculated as cluster frequency (4.55), thus number of elements in a cluster on total number of elements, while seasonal probability (4.54) is calculated as already described for each typical day of the *centroids*. Then, yearly probability is multiplied for seasonal one and divided for the sum of all obtained probabilities, in order to get one once summed all of them (4.56).

$$p_{seasonal,tot,k} = p_{day,winter,k} * p_{day,spring,k} * p_{day,summer,k} * p_{day,autumn,k} \quad \forall k \in centroids \tag{4.54}$$

$$p_{yearly,k} = \frac{n_{elements,k}}{n_{elements,tot}} \quad \forall k \in centroids \tag{4.55}$$

$$p_{final,k} = p_{seasonal,tot,k} * p_{yearly,k} / \sum p_{seasonal,tot,k} * p_{yearly,k} \; \forall k \in centroids \tag{4.56}$$

Even if this process is presented for transparency, its solutions are not presented in this work. Indeed, results are very unprecise, especially with k=2, which gives strongly unrealistic ones, even though it should be the best case, with the highest ASV. The point is that a high quality of clusters does not entail a high quality solution. Hence, it is necessary to start from the solution to decide the appropriate number of clusters.

4.5 Robustness test

4.5.1 Operation of the residential MES tested on years from 2010 to 2020 in a row

This test provides indications about the quality of a given solution. Indeed, the deterministic (or stochastic) model for the residential MES, described in Section 4.1 (Section 4.3), gives the optimized decision variables for the first-stage (design) and the second-stage (operation). However, the optimization period is of just few time steps and, in addition, representative days are generated starting from the period 2005-2009. The idea is to be at 2010 and try to "predict" how solutions, obtained with data from the past, will work in future, with respect to the reference model for that period (Section 4.1).

Hence, the robustness test allows seeing that: differently from the previous models, the design decision variables are set as input data. Other inputs, such as global solar irradiance, air temperature, electricity price and energy demand are taken from historical data, implying 11*365*24=93690 time steps. There is no necessity of generating representative days, because of the reduced computational complexity given by the set design variables. In fact, this test is an optimization for the operation, all along the testing dataset.

Some changes are done in the objective function and in constraints for the energy conversion system.

$$min\, f = min\left(Cost_{des} + \sum_{y=y_0}^{Y} \sum_{d=0}^{D} \sum_{h=0}^{H} Cost_{op,t} \right) \quad (4.57)$$

The total cost has the same expression as the one indicated in Section 4.1, but the design part is a parameter, because all the sizes are already set.

For what concerns the PV system, from equation (4.8), the design decision variable is the area of the system. Now, this value is set, as well as the other parameters are. Consequently, the energy flux P_{PV} is now a parameter too, instead of being a dependent variable.

Equations for the EES are the same as the one already presented, (4.11-18). Same for the thermal energy storage (equations 4.35-42), the grid (equations 4.43, 4.44) and the energy balances (4.45, 4.46).

Regarding CHP, equations (4.19), (4.20) and (4.27) are still valid. However, the additional variable $\vartheta_{\mathrm{CHP,t}}$ is not required anymore, because size of the system is set, avoiding bilinear constraints. To simplify the model, equations (4.22-26) become:

$$P_{CHP,t} \leq \delta_{CHP,t} * C_{CHP} \quad (4.58)$$

Which is possible because C_{CHP} is not a decision variable anymore. Additionally, in equation (4.27) $\delta_{CHP,inv}$ is a parameter.

Similarly, the boiler is managed as the CHP. Indeed, the additional variable $\vartheta_{\mathrm{boiler,t}}$ is not required anymore and equations (4.29-33) become:

$$P_{boiler,t} \leq \delta_{boiler,t} * C_{boiler} \quad (4.59)$$

4.5.2 Operation of the residential MES optimized separately on each year from 2010 to 2020

This test is done to see how good a solution is with respect to a single yearly reference one. Equations, constraints, input parameters and output variables are the same as for the other test in Section 4.4.1, but operation lasts just one year. Hence, 11 models, because of the length of the testing dataset, are developed, with 8760 time steps per model. The computational effort is not too high because the design variables are set. Note that the design cost has to be taken back to a single year.

5. Results and discussion

The following Section collects the results for the different optimization processes. In particular, Section 5.1 studies the influence of the air temperature on the clustering process, comparing generation of representative days considering just the global solar irradiance as uncertain parameter or considering air temperature too; Section 5.2 presents results, concerning total cost and design decision variables, for the reference model of the residential MES for the period 2010-2020 and for reference annual models; Section 5.3 analyses optimised solutions for the deterministic model with N representative days, comparing the methods to generate them with respect to the training dataset; Section 5.4 presents results for the two-stage stochastic model with N representative days and M scenarios, regarding the training dataset; Section 5.5 compares objective functions of the previous models, to understand which one is closer to the reference solution for the testing dataset.

5.1 Influence of air temperature on clustering process

As already explained in Section 4.3, global solar irradiance and air temperature are the two attributes for the clustering process. The idea is to keep seasons as separated as possible, without rigid distinctions, to represent the initial dataset in the best possible way. In particular:

1. The introduction of air temperature should allow to assign elements in a better way. Indeed, with only global solar irradiance as uncertain parameter, a cluster with a representative day from summer could include an element from winter. In this case, it would be represented by the irradiance, but other parameters, such as thermal demand, would not be representative for it.
2. The division of the dataset into seasons should decrease the quality of the solution. In fact, if a representative element of a cluster is “between” two seasons, it would be well represented by half elements from one season and half from the following one.

To demonstrate these two points, let us focus on a single simulation. Please note that every time the clustering process is repeated, results change. In particular, as seen in Section 4.4.2, representative days are not the same, as well as number of days per seasons are not. Table 5.1.2 shows that, if air temperature is considered as attribute, number of representative days per season are different with respect to the case with only global solar irradiance. In that case, N is set arbitrarily equal to 12. This result is useful to validate the methodology for the two-stage stochastic model for the residential MES, presented in Section 4.4.

In any case, table 5.1.1 presents results for a single clustering process, repeated with air temperature and global solar irradiance as attributes for the k-means process (left side) or with just global solar irradiance (right side). For each side, the series of representative days is presented. In particular, focusing on a single day, the “title” is its date and, knowing it, the corresponding season can be considered. For instance, the first representative day of the left side is the 10th of March 2007, which is a winter day. Then, the season of each element of the cluster is found, to understand which is the percentage of well-positioned elements, meaning an element belonging to the same season of the representative element of its belonging cluster. For the previous example, 112 elements belong to that cluster: 59 of them are winter days, 47 are spring days, 1 is a summer day and 5 are autumn days. Considering the average value of well-positioned elements for all clusters, the inclusion of air temperature implies a better assignment of the elements, with 59,65% of well-positioned ones, instead of 50,39% with only global solar irradiance. Additionally, “border” clusters should not be considered in this analysis. For instance, the 10th of March is formally winter, but has irradiance and temperature profiles similar to spring days too. Hence, avoiding these clusters, the percentage passes from 51,11%, without considering air temperature, to 62,18% including it.

To deepen point 1, let us consider a representative day strongly linked to climate conditions of a season. For instance, figure 5.1.1 represents profiles for the 3rd of February 2007. Considering table 5.1.2, this day is a representative one just without air temperature as uncertain parameter. It shows that 12 elements are from summer season, which is reliable focusing on the irradiance, possibly related to a cloudy day. However, if the air temperature is included in the analysis, it appears obvious that such elements are bad-positioned, because the temperature profile is not coherent with a summer day in Padova. Same consideration could be done with the thermal energy demand, which strongly depends on the external temperature. Therefore, air temperature seems necessary to get high quality solutions.

In addition, it is possible to compare results for the 20th of May 2006 (figure 5.1.3), which is a representative day in both simulations. Comparing left and right side of the table, the improvement of introducing air temperature as uncertain parameter is evident, because a higher number of elements belongs to the correct season, spring, passing from 60,49 to 65,17%. In both cases autumn and winter days are not included in this cluster, but there is an improvement of the seasonal assigment, which means a better representation of the dataset.

With respect to point 2, let us consider the case of 28th of September, as shown in figure 5.1.2, common for both sides of table 5.1.1. If temperature is included in the clustering process, most of the days belonging to that cluster are summer or autumn ones, which is coherent to the profile indicated in figure 5.1.2. Note that no elements are from winter, which is a great result, because it is improbable to 25°C in winter season. On the contrary, if the air temperature is not included, 33,88% of elements do not belong to summer or autumn. In particular, 11,11% of elements belong to winter, which means they are strongly bad-positioned. The inclusion of air temperature seems to be relevantly positive.

With temperature			**Without temperature**		
	10/03/2007			**03/02/2007**	
Winter	59 days	52,68 %	**Winter**	142 days	42,26 %
Spring	47 days	41,96 %	**Spring**	37 days	11,01 %
Summer	1 day	0,89 %	**Summer**	12 days	3,57 %
Autumn	5 days	4,46 %	**Autumn**	145 days	43,15 %
	02/04/2009			**12/04/2009**	
Winter	10 days	5,49 %	**Winter**	31 days	40,26 %
Spring	54 days	29,67 %	**Spring**	25 days	32,47 %
Summer	12 days	6,59 %	**Summer**	19 days	24,68 %
Autumn	106 days	58,24 %	**Autumn**	2 days	2,60 %
	28/08/2006			**18/01/2008**	
Winter	0 days	0 %	**Winter**	62 days	36,90 %
Spring	43 days	43 %	**Spring**	3 days	1,79 %
Summer	49 days	49 %	**Summer**	1 days	0,60 %
Autumn	8 days	8 %	**Autumn**	102 days	60,71 %
	29/10/2007			**24/06/2008**	
Winter	14 days	9,86 %	**Winter**	0 days	0,00 %
Spring	34 days	23,94 %	**Spring**	73 days	52,14 %
Summer	1 day	0,70 %	**Summer**	67 days	47,86 %
Autumn	93 days	65,49 %	**Autumn**	0 days	0,00 %
	30/08/2008			**18/04/2006**	
Winter	0 days	0,00 %	**Winter**	4 days	4,30 %
Spring	3 days	1,94 %	**Spring**	37 days	39,78 %
Summer	152 days	98,06 %	**Summer**	33 days	35,48 %
Autumn	0 days	0,00 %	**Autumn**	19 days	20,43 %

	13/01/2007			28/04/2009	
Winter	135 days	64,59 %	**Winter**	5 days	6,85 %
Spring	6 days	2,87 %	**Spring**	41 days	56,16 %
Summer	0 days	0,00 %	**Summer**	27 days	36,99 %
Autumn	68 days	32,54 %	**Autumn**	0 days	0,00 %
	14/01/2007			**08/07/2006**	
Winter	83 days	64,84 %	**Winter**	0 days	0,00 %
Spring	8 days	6,25 %	**Spring**	34 days	24,46 %
Summer	0 days	0,00 %	**Summer**	105 days	75,54 %
Autumn	37 days	28,91 %	**Autumn**	0 days	0,00 %
	28/09/2006			**28/09/2006**	
Winter	0 days	0 %	**Winter**	20 days	11,11 %
Spring	9 days	6,25 %	**Spring**	41 days	22,77 %
Summer	60 days	41,66 %	**Summer**	40 days	22,22 %
Autumn	75 days	52,08 %	**Autumn**	79 days	43,88 %
	06/05/2006			**09/02/2007**	
Winter	0 days	0,00 %	**Winter**	80 days	42,11 %
Spring	127 days	90,71 %	**Spring**	31 days	16,32 %
Summer	13 days	9,29 %	**Summer**	12 days	6,32 %
Autumn	0 days	0,00 %	**Autumn**	67 days	35,26 %
	18/07/2006			**21/08/2009**	
Winter	0 days	0,00 %	**Winter**	0 days	0,00 %
Spring	71 days	31,98 %	**Spring**	109 days	47,81 %
Summer	151 days	68,02 %	**Summer**	119 days	52,19 %
Autumn	0 days	0,00 %	**Autumn**	0 days	0,00 %
	20/11/2005			**22/02/2009**	
Winter	149 days	73,76 %	**Winter**	93 days	73,2 %
Spring	0 days	0,00 %	**Spring**	0 days	0,00 %
Summer	0 days	0,00 %	**Summer**	3 days	2,40 %
Autumn	53 days	26,24 %	**Autumn**	31 days	24,4 %
	20/05/2006			**20/05/2006**	
Winter	0 days	0,00 %	**Winter**	0 days	0,00 %
Spring	58 days	65,17 %	**Spring**	49 days	60,49 %
Summer	31 days	34,83 %	**Summer**	32 days	39,51 %
Autumn	0 days	0,00 %	**Autumn**	0 days	0,00 %

Table 5.1.1: Days per season belonging to a representative day. In particular, days on the left consider the clustering process with air temperature and global solar irradiance as uncertain parameters, while days on the left consider only global solar irradiance. For each day, its date is indicated and all the elements belonging to it, divided per season. Then, on the right column per cluster, the percentage of days belonging to the correct season is indicated.

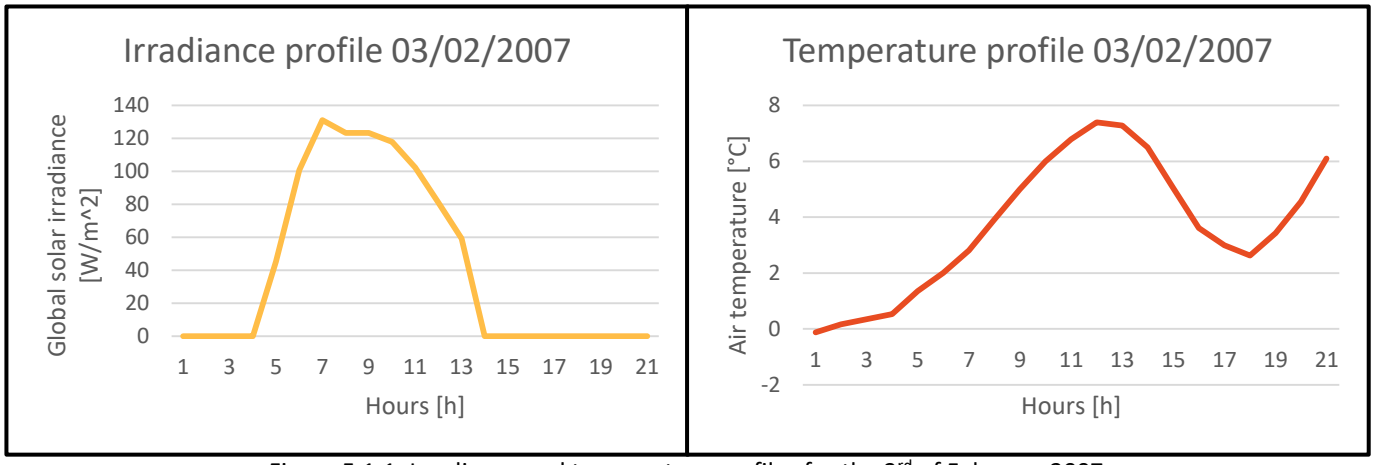

Figure 5.1.1: Irradiance and temperature profiles for the 3rd of February 2007.

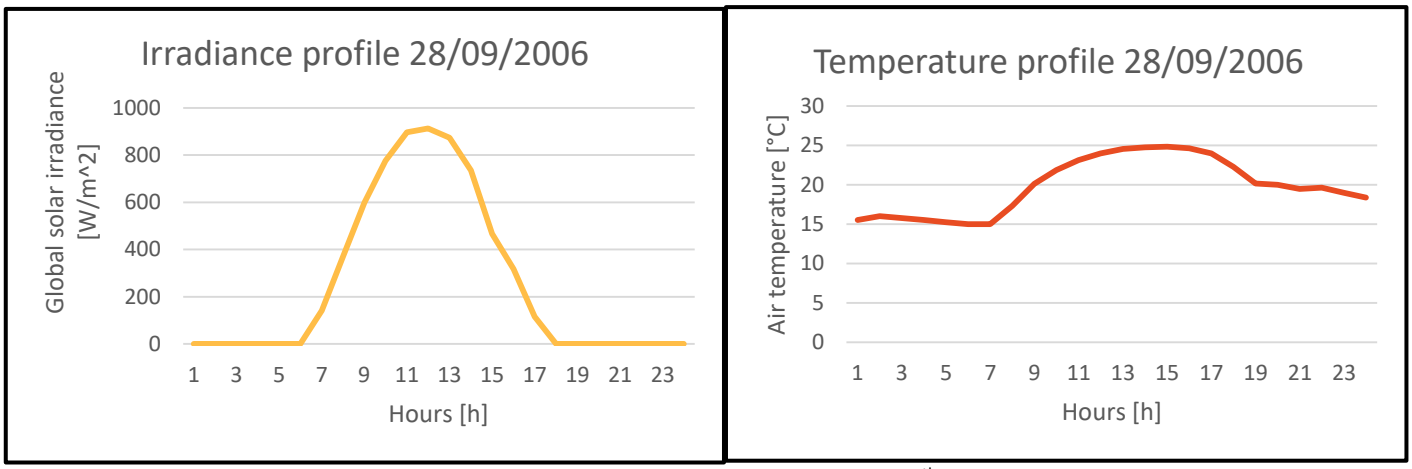

Figure 5.1.2: Irradiance and temperature profiles for the 28th of February 2006.

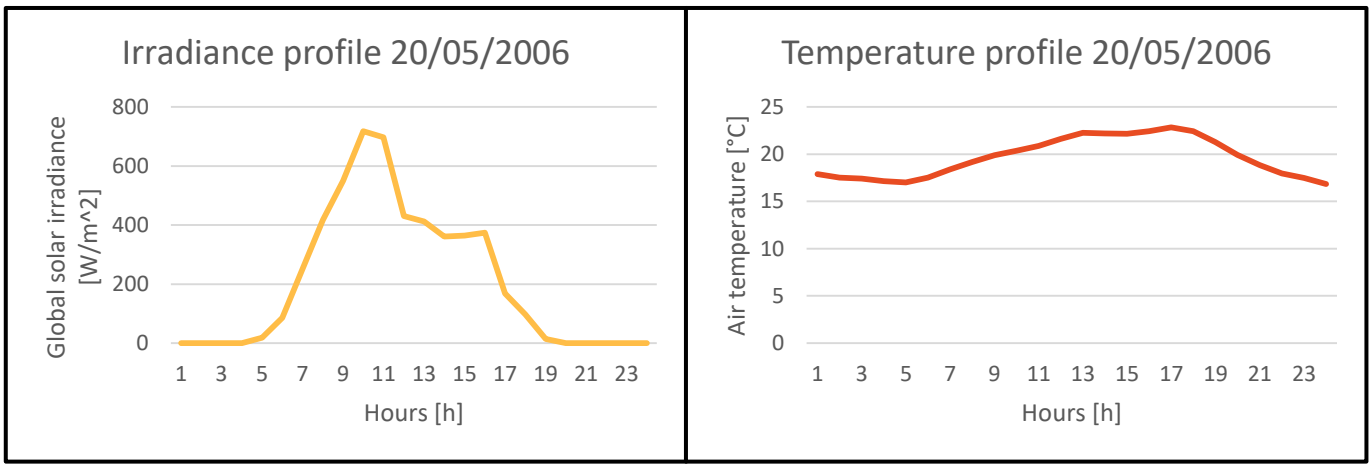

Figure 5.1.3: Irradiance and temperature profiles for the 20th of March 2006.

		Representative days	
		With air temperature	Without air temperature
Simulation 1	Winter	3	5
	Spring	2	3
	Summer	3	3
	Autumn	4	1
Simulation 2	Winter	4	5
	Spring	2	3
	Summer	3	3
	Autumn	3	1
Simulation 3	Winter	3	3
	Spring	2	3
	Summer	3	3
	Autumn	4	3
Simulation 4	Winter	3	3
	Spring	2	3
	Summer	3	3
	Autumn	4	3
Simulation 5	Winter	3	3
	Spring	2	3
	Summer	3	3
	Autumn	4	3
Simulation 6	Winter	3	4
	Spring	2	3
	Summer	3	3
	Autumn	4	2
Simulation 7	Winter	3	4
	Spring	2	3
	Summer	3	4
	Autumn	4	1
Simulation 8	Winter	3	3
	Spring	2	1
	Summer	3	5
	Autumn	4	3
Simulation 9	Winter	3	2
	Spring	3	3
	Summer	3	3
	Autumn	3	4
Simulation 10	Winter	3	2
	Spring	2	3
	Summer	3	3
	Autumn	4	4

Table 5.1.2: Number of representative days per season by repeating the clustering process. Additionally, the comparison is done considering the air temperature and the global solar irradiance as uncertain parameters (on the left) or just considering the global solar irradiance (on the right).

5.2 Reference models of the residential MES: objective function and design decision variables

5.2.1 Optimization results on a single year

As already explained in Section 4, all the following models are referred to the residential MES scheme of Section 2. In particular, the ones in this sub-Section are solved for the respective year, implying 8760 time steps: they do not require time series aggregation.

Then, the ones referred to years belonging to the training dataset (2005-2009) are used as reference to get the best number of representative days, based on MAE and RMSD (Section 4.3.2). On the contrary, the ones for the testing dataset (2010-2020) are used as reference to test the quality of the proposed solution and, consequently, the method to obtain it, as explained in Section 4.4.2.

	C_TOT	C_DES	C_OP	C_PV	C_CHP	C_boiler	E_EES	Q_TES
	k€	k€	k€	kWp	kW	kW	kWh	kWh
2005	59,85	36,98	22,87	224,52	74,13	73,65	0,00	219,83
2006	61,72	37,63	24,09	164,90	78,44	67,19	85,99	207,89
2007	66,64	28,95	37,70	145,44	55,22	106,09	18,54	222,40
2008	62,15	32,60	29,55	119,29	76,96	70,96	76,82	204,09
2009	65,89	33,76	32,14	167,45	59,55	101,91	67,63	170,67
2010	61,41	35,61	25,81	166,73	79,12	68,53	62,69	164,84
2011	63,12	45,82	17,31	249,64	75,49	73,05	62,22	335,06
2012	61,54	34,36	27,18	127,86	79,00	64,40	62,92	295,56
2013	69,84	50,41	19,43	219,58	79,42	66,71	163,52	353,16
2014	64,59	45,57	19,02	225,62	94,09	43,45	99,61	184,70
2015	61,93	45,36	16,57	248,47	78,20	72,81	100,37	158,61
2016	62,69	39,63	23,05	178,18	80,66	72,55	99,59	185,06
2017	59,62	30,82	28,80	125,27	78,26	68,71	41,73	191,98
2018	66,43	44,32	22,11	129,48	133,40	49,87	68,12	300,72
2019	67,40	51,91	15,49	250,19	95,03	42,89	158,44	193,11
2020	61,65	48,60	13,06	280,85	80,02	65,67	80,91	232,17

Table 5.2.1: Costs and design variables for the yearly reference solutions. With respect to the reference model for each year of the training and testing dataset, the first three columns indicate the total cost for the year, the design cost and the operation one. The following columns indicate the size of the PV system, the CHP, the boiler, the electrical storage and the thermal one.

With respect to table 5.2.1, it seems clear that sizes for each unit are different year by year, depending on the global solar irradiance and the air temperature. In fact, prices of energy carriers, energy demands and techno-economic characteristics for the energy system components are the same. In particular, year 2005 presents the lowest cost, related to the high size of the PV system and the absence of the electric storage. This is possible because of good conditions regarding photovoltaic production. However, it is difficult to comment how the sizes vary year by year, because of the combination of many decision variables in the optimization problem.

5.2.2 Optimization results for the period 2010-2020

The reference model for the testing dataset solves the problem for the "future" (2010-2020), setting the lower bound for the total cost. Considering eleven years in a row, number of time steps is equal to 96360, forcing to use clustering techniques, as already indicated in Sections 4.1.

The number of representative days N varies from 2 to 45, obtaining N-1 optimized solutions (design and operation decision variables) from N-1 sets of representative days. Then, design variables are set and, because of the reduction of variables, the operation is optimized on the entire testing dataset. The tested solution with the lowest objective function is the closest to the actual solution of the model, even if it is not.

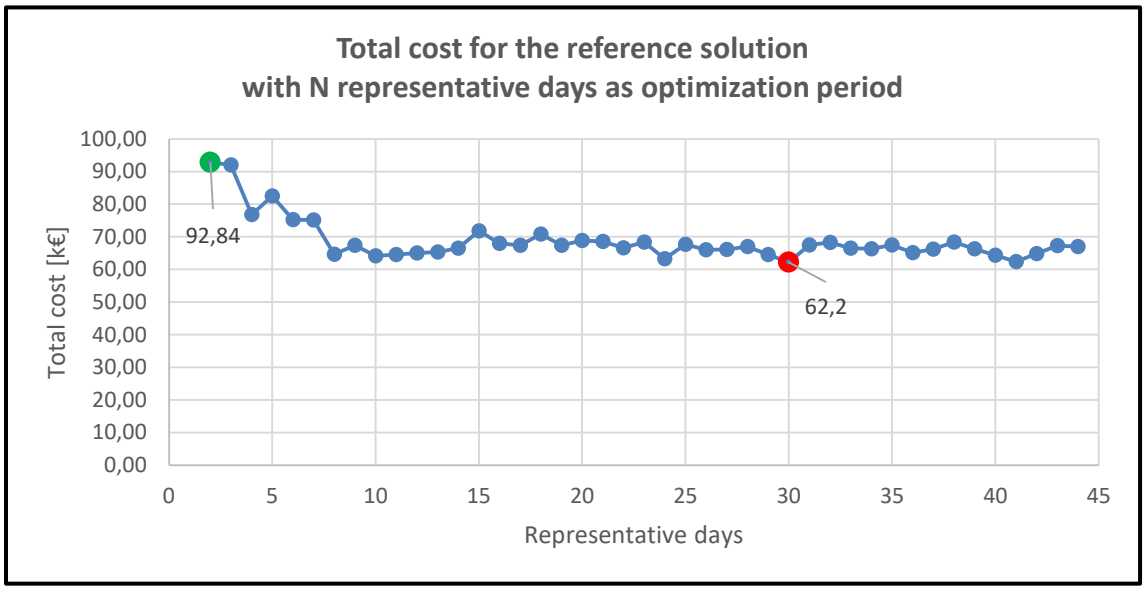

Figure 5.2.1: Total cost for the reference solution with N representative days as optimization period. The total cost [k€], on the y-axis, taken back to one single year, is function of the number of representative days used as input for the model, on the x-axis. The model finds the minimum cost by evaluating design and operation decision variables. Typical days are obtained from dataset 2010-2020. The green point indicates the maximum value, obtained for 2 representative days, while the red one represents the minimum one, for N=30.

The idea to test the solution is related to the necessity of obtaining real costs. In fact, by considering just a limited number of representative days, the risk is to overestimate (or underestimate) them. To understand that, let us consider figure 5.2.1, which represents the profile of the objective function solving the reference model with N representative days as optimization period. For N=2 (green point), the objective function is strongly higher with respect to the minimum, at N=30 (red point). This aspect is linked to the addition of the worst thermal energy demand scenario. As explained in Section 4, it is not representative, but it is necessary to size properly the components. Hence, if this scenario is forced to be representative for a cluster, the overestimation for the objective function is relevant.

The trend is coherent, according to clustering theory [54-55]: if the technique was k-means, the trend should be monotonic, due to the use of centroids as representative elements, which means there is always an underestimation of the total cost. On the contrary, k-medoids [55] does not entail such a trend, but an oscillating one. Therefore, the only difference is given by the necessary overestimation, due to the introduction of thermal demand extreme scenario.

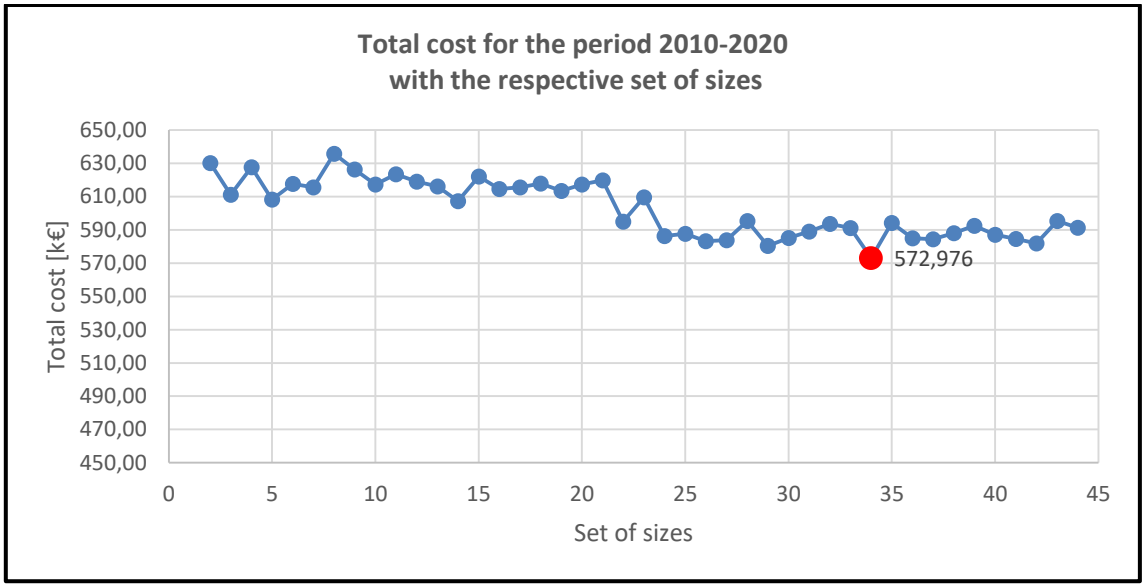

Figure 5.2.2: Total cost for the period 2010-2020 with the respective set of sizes. The objective function, on y-axis, for the entire period 2010-2020, is indicated as function of the respective set of sizes. This cost is obtained optimizing the operation, which means the design decision variables are set solving the model with N representative days, while here only operation variables are free to vary. The minimum corresponds to the set of sizes related to 34 days as optimization period.

Figure 5.2.2 indicates the actual cost, obtained testing the set of sizes for the entire testing dataset, optimizing the operation. Note that, for set of sizes, it is meant an array as $[C_{PV}, C_{CHP}, C_{boiler}, C_{EES}, C_{TES}]$, as indicated in table 6.2.1. The trend of figure 5.2.2 is different from the one of figure 5.2.1, because input data are historical one, without data aggregation.

The minimum in figure 6.2.2 is obtained for the set of sizes corresponding to 34 representative days, while in figure 6.2.1 30 days give the lowest cost. Moreover, the maximum cost in figure 6.2.2 is not linked to sizes obtained with N=2 representative days, which proves the effect of overestimation of the total cost related to the clustering process. Therefore, the reference solution is the set of sizes corresponding to an optimization period of N=34 days.

Table 5.2.1 collects the total cost for the entire period 2010-2020, the one for the optimization period corresponding to N representative days and the set of sizes for each set of representative days. Please note that, for low values of N, the boiler is absent, while the CHP is overestimated: this is related to the independence between the representative days, which becomes less important once the number of clusters increases. Furthermore, the use of the EES, as well as an underestimation of the size for the PV system, entails the highest values for the total cost. For example, if N=12 or 14, EES is included, so the cost is strongly higher than for N=13, when EES is not present in the optimized set of sizes.

	C_TOT	C_TOT_N	C_PV	C_CHP	C_boiler	E_ EES	Q_ TES
	k€	k€	kWp	kW	kW	kWh	kWh
2	630,11	92,84	41,52	123,37	0,00	0,00	101,43
3	611,15	91,99	98,98	123,05	0,00	0,00	135,01
4	627,69	76,78	70,09	122,64	0,00	58,00	135,43
5	608,21	82,49	114,04	122,35	0,00	0,00	149,34
6	617,63	75,19	88,30	121,69	0,00	0,00	181,03
7	615,49	75,17	79,83	122,60	0,00	0,00	137,21
8	635,71	64,63	43,43	122,55	0,00	48,01	139,70
9	626,23	67,42	60,61	122,20	0,00	34,18	156,28
10	617,25	64,13	104,86	122,25	0,00	41,60	153,91
11	623,44	64,53	77,36	121,87	0,00	35,70	172,09
12	619,03	65,00	100,55	121,81	0,00	37,84	174,96
13	616,20	65,31	124,54	121,98	0,00	43,72	167,14
14	607,33	66,54	140,46	121,99	0,00	0,00	166,40
15	622,12	71,81	73,62	122,02	0,00	29,88	165,27
16	614,51	67,94	100,22	121,79	0,00	0,00	175,95
17	615,54	67,35	97,31	121,70	0,00	0,00	180,63
18	617,94	70,83	77,55	122,29	0,00	0,00	152,27
19	613,40	67,40	102,85	121,94	0,00	0,00	169,11
20	617,24	68,81	88,12	121,77	0,00	0,00	177,06
21	619,73	68,57	84,07	121,48	0,00	0,00	191,05
22	594,86	66,64	95,14	102,46	35,52	0,00	183,71
23	609,51	68,35	142,92	121,48	0,00	0,00	191,05
24	586,26	63,22	140,46	102,69	35,32	0,00	174,72
25	587,57	67,66	126,26	105,68	37,12	0,00	135,74
26	583,17	66,02	124,35	97,86	50,17	0,00	162,76
27	583,87	66,09	149,19	103,10	39,48	0,00	158,59
28	595,36	66,96	121,46	101,30	36,93	0,00	228,70
29	580,28	64,53	113,11	95,33	51,80	0,00	130,89
30	585,09	62,2	131,66	102,28	37,045	0,00	157,57
31	588,96	67,46	118,17	104,03	36,78	0,00	157,57
32	593,61	68,30	109,49	102,10	37,03	0,00	197,51
33	591,02	66,50	111,72	101,20	41,51	0,00	185,50
34	572,97	66,33	312,00	106,82	33,16	0,00	149,57
35	594,21	67,48	124,35	100,56	36,26	0,00	228,69
36	584,93	65,15	131,66	102,07	37,42	0,00	157,10
37	584,43	66,16	111,72	98,67	46,02	0,00	150,39
38	588,02	68,36	98,91	99,37	44,53	0,00	157,72
39	592,40	66,27	112,31	98,53	45,12	0,00	215,49
40	587,10	64,30	99,56	94,62	50,12	0,00	182,43
41	584,55	62,36	124,41	97,10	44,97	0,00	181,27
42	581,78	64,86	173,48	102,84	35,41	0,00	168,62
43	595,28	67,26	111,72	99,58	42,32	0,00	226,73
44	591,24	67,03	102,01	98,20	42,85	0,00	196,52

Table 5.2.1: Table including costs and design decision variables for the reference model with N representative days. Columns indicate: total cost for the tested solution on the entire period 2010-2020, total cost for the solution with N representative days, size of the photovoltaic system, size of the CHP system, size of the boiler, size of the electric energy system, size of the thermal energy system.

5.3 Deterministic model of the residential MES with N representative days, for the period 2005-2009: objective function and decision variables

5.3.1 Optimization results with representative days from annual dataset

This section includes results obtained with the deterministic model with N representative days, whose generation is carried out through clustering techniques, starting from annual dataset (Section 4.3.2). The reference models are the annual ones (results in Section 5.2). The goal is to obtain the number of clusters that minimizes the MAE and the RMSD, as explained in Section 4.3.2.

To summarize, N representative days are generated for each year of the training dataset, with $N \in [2,30]$. The deterministic model of the residential MES is solved for each set of clusters and the optimized objective function is compared to the reference one for the respective year, evaluating the Mean Absolute Error and the Root Mean Square Difference.

Table 5.3.1 collects all this data. The best solution is to adopt a number of representative days equal to 27, which minimizes both the MAE and the RMSD. The actual days are taken from year 2007, which presents the minimum value of these two indicators: this set of representative days will be used to test the quality of the solution for the training and testing dataset.

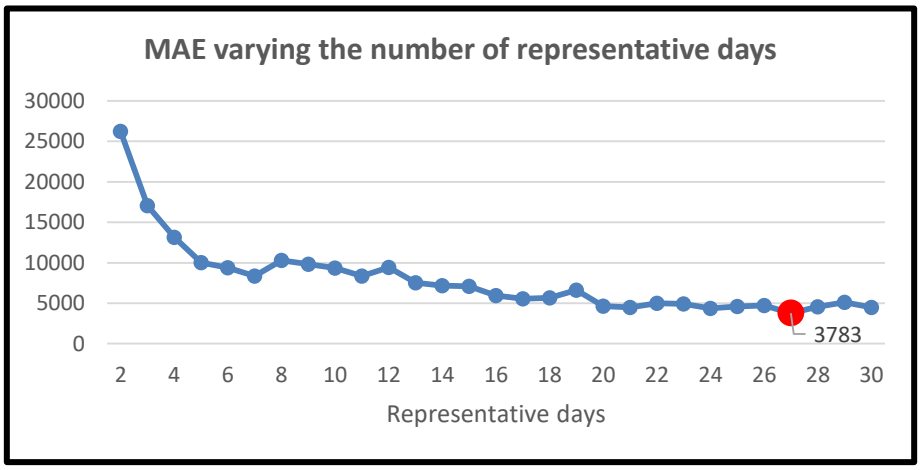

Figure 6.3.1: Mean Absolute Error varying the number of representative days. MAE, on the y-axis, calculated for the total cost, with respect to the reference yearly solution, for different number of clusters, on the x-axis.

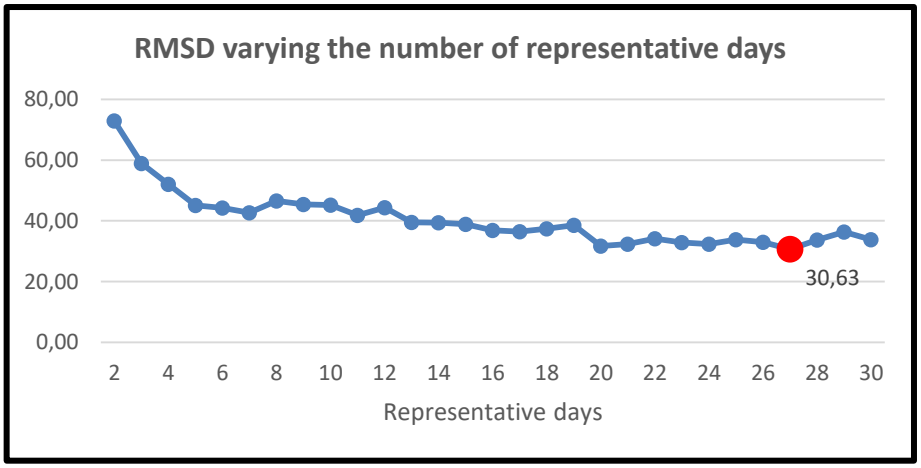

Figure 6.3.2: Root Mean Square Difference varying the number of representative days. RMSD, on the y-axis, calculated for the total cost, with respect to the reference yearly solution, for different number of clusters, on the x-axis

Year	4	3	2	1	0	MAE	RMSD
Reference	2009 [€]	2008 [€]	2007 [€]	2006 [€]	2005 [€]		
2	65892	62147	66644	61719	59848		
3	92003	92230	89307	82087	91858	26247	72,95
4	81702	75108	87156	82809	74882	17081	58,95
5	82585	72939	79111	70451	76964	13160	52,08
6	73129	74185	78269	71161	69698	10038	45,13
7	71111	75313	77675	69463	69693	9401	44,25
8	70252	77045	72656	70736	67430	8374	42,71
9	71126	77989	76956	70576	71074	10294	46,59
10	70270	72578	75620	75000	71888	9821	45,38
11	71002	74799	70224	75043	71874	9339	45,19
12	70810	72950	74052	68520	71700	8356	41,82
13	70363	70355	77200	73972	71515	9431	44,38
14	70611	69994	73101	69130	70990	7515	39,51
15	71652	72521	68644	69715	69508	7158	39,43
16	71847	70335	69449	69384	70640	7081	38,88
17	71563	68780	66529	69373	69602	5965	36,83
18	66684	71277	68151	70115	67797	5555	36,42
19	67380	72548	66870	68824	69032	5681	37,43
20	71610	71558	65985	68992	69962	6635	38,58
21	68389	67494	69212	67376	67068	4658	31,66
22	66985	68127	68056	68151	67356	4485	32,34
23	68932	70879	66409	67655	66848	4989	34,14
24	69614	69067	65549	67255	67096	4904	32,88
25	68100	70192	66808	66174	66734	4351	32,35
26	65273	70951	65877	67247	67236	4621	33,82
27	69620	70066	65914	65398	67425	4727	32,99
28	67159	65429	67819	66158	68601	3783	30,63
29	62498	69269	66650	64336	69437	4546	33,69
30	68788	72723	66987	63824	69512	5117	36,35
	65569	70296	67199	66576	68384	4484	33,81

Table 5.3.1: MAE and RMSD, varying the number of representative days, for the annual clustering method. Columns indicate the objective function obtained as solutions of the deterministic models with N representative days, with respect to a specific year of the training dataset. The row “reference” indicates the reference model for the respective year. Columns MAE and RMSD indicate the Mean Absolute Error and the Root Mean Square Difference.

An interesting result is shown in table 5.3.1: the choice is to take the days from year 2007, even if 2008 presents a lower value for the objective function. This is due to the fact that the solution for 2008 is cheaper (65429€) compared to 2007 (67819€), which means that the solution for 2008 with 27 representative days is far from the reference value. Indeed, every year can be different, depending on weather condition, entailing naturally lower costs.

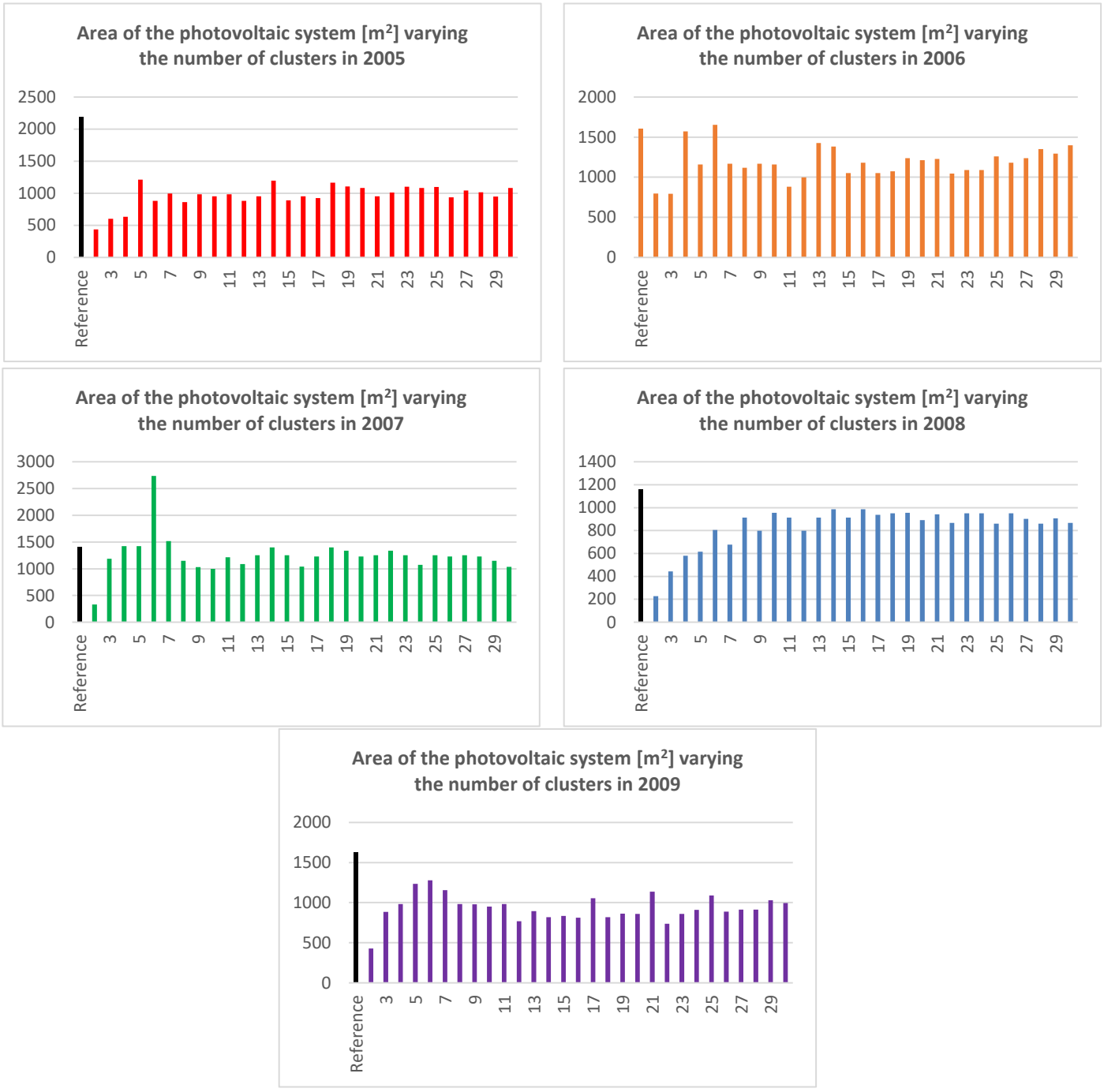

Figure 5.3.3: Area of the photovoltaic system [m^2] obtained from the different deterministic models with N representative days, for different years.

As shown in figure 5.3.3, the area of the PV system is usually underestimated with respect to the annual reference model. This is linked to the clustering process because, as already deepened, the introduction of extreme scenarios causes disequilibrium in the representation of the dataset. The underestimation is particularly relevant for a low number of representative days, while it tends to stabilize while increasing N, excluding atypical values. In any case, it continues oscillating, as the objective function does.

Considering now figure 5.3.4, it appears clear the difference in the CHP size for the models, independently on the number of days, with respect to the reference yearly one: this is related to the independence of the different scenarios, forcing to get a set level of energy in the storage systems at midnight. Therefore, the reference model is able to exploit the link among days, differently from these models with N days, whose rigidity causes an overestimation of the CHP size, as well as of the total cost. Similarly, because of the high

size of the CHP, the boiler is usually not included, unless for high number of representative days and, in any case, sporadically.

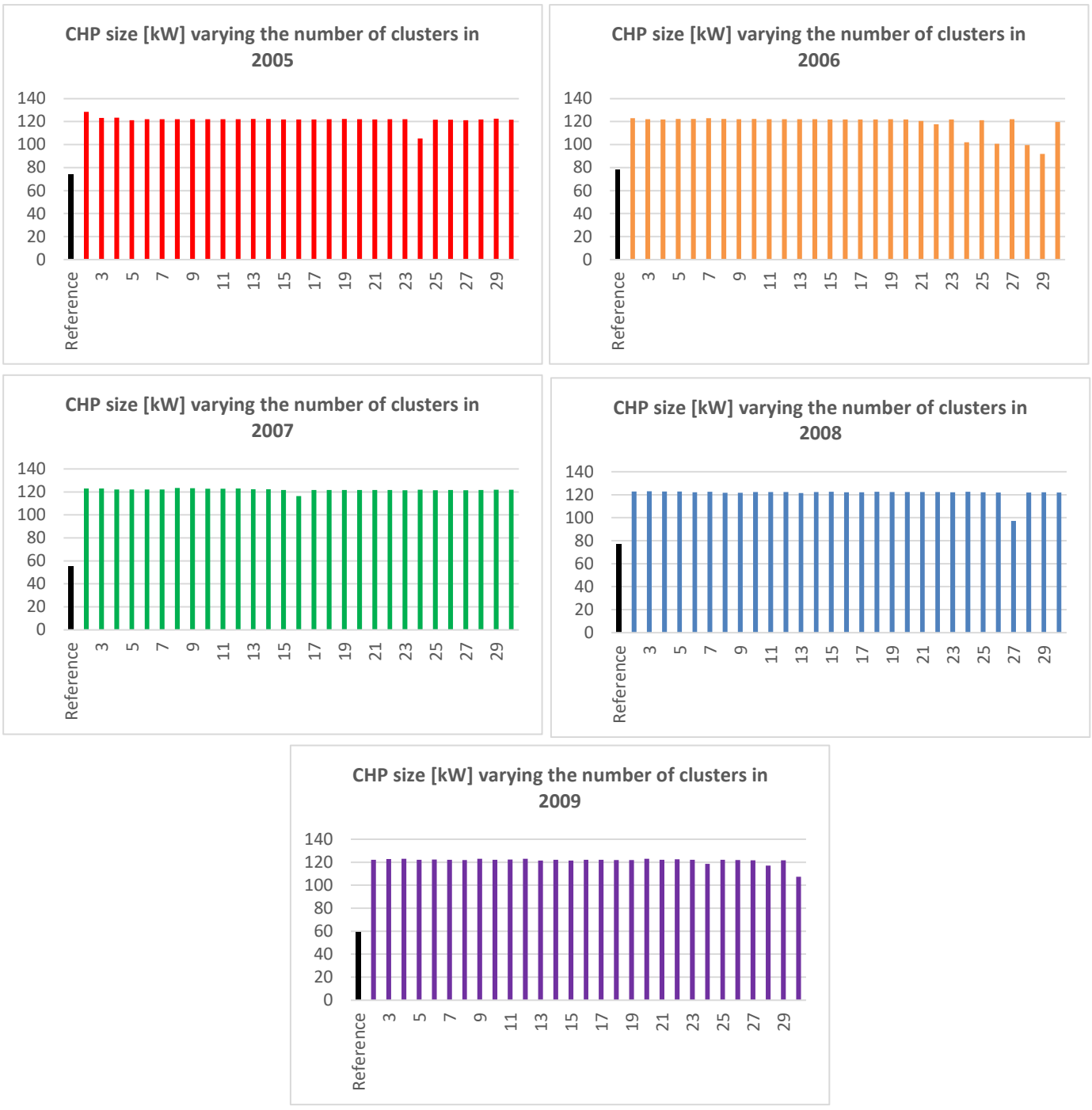

Figure 5.3.4: Size of the CHP system [kW] obtained from the different deterministic models with N representative days, for different years.

5.3.2. Optimization results with representative days from the whole training dataset

In this case, the procedure is the same as the one presented in Section 4.3.2: sets of representative days are obtained varying the number N of clusters, thus through clustering process for the entire training dataset, 2005-2009. Then, the deterministic model is solved to obtain optimized design and operation decision variables (figure 5.3.5). Finally, design variables are set and the solution is tested for the entire training dataset, optimizing the operation for five years (figure 5.3.6).

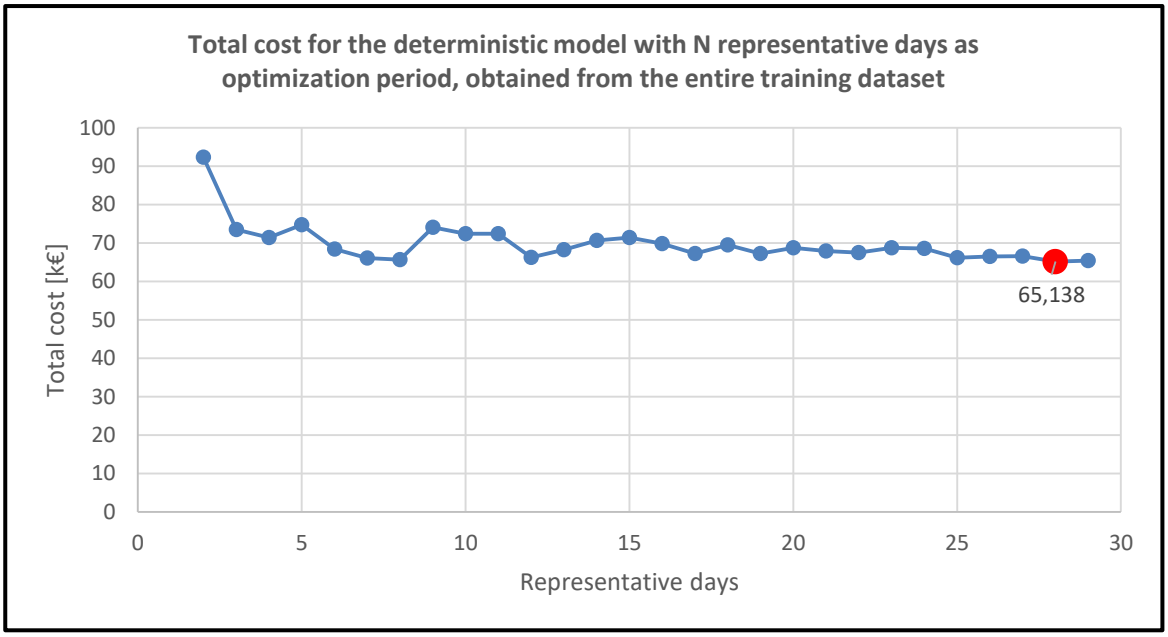

Figure 5.3.6: Total cost for the deterministic model with N representative days as optimization period, obtained from the entire training dataset. The objective function is optimized starting from the generated sets of clusters. The red point is the minimum value of total cost.

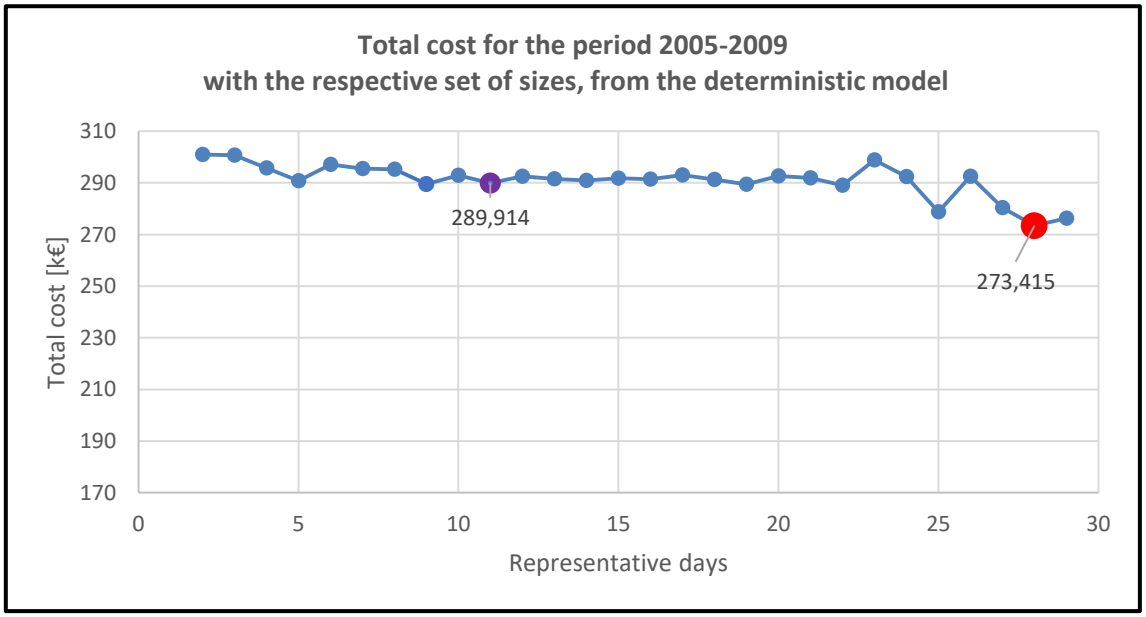

Figure 5.3.7: Total cost for the period 2005-2009 with the respective set of sizes, from the deterministic model. The objective function is obtained optimizing the operation for the training dataset. This means the design decision variables are set, once solved the deterministic model with N representative days. Red point is the minimum one, while purple ones can be used for the comparison with the two-stage stochastic model.

	C_TOT	*C_TOT_N*	*C_PV*	*C_CHP*	*C_boiler*	*E_EES*	*Q_TES*
	k€	**k€**	**kWp**	**kW**	**kW**	**kWh**	**kWh**
2	301,00	92,29	38,50	123,05	0,00	23,07	129,72
3	300,70	73,54	55,14	122,13	0,00	26,61	159,74
4	295,75	71,38	67,59	122,42	0,00	35,31	145,81
5	290,80	74,74	126,29	121,94	0,00	0,00	169,07
6	297,08	68,40	205,26	122,44	0,00	59,74	145,08
7	295,51	66,04	126,08	120,38	0,00	0,00	246,77
8	295,27	65,62	157,14	121,74	0,00	40,50	178,68
9	289,51	74,09	157,14	122,06	0,00	0,00	163,32
10	292,93	72,38	93,79	121,52	0,00	0,00	189,23
11	289,91	72,43	121,12	122,44	0,00	0,00	145,05
12	292,56	66,24	118,95	121,33	0,00	0,00	198,59
13	291,51	68,21	101,36	122,02	0,00	0,00	165,08
14	290,92	70,63	106,53	121,92	0,00	0,00	169,71
15	291,75	71,43	98,20	121,92	0,00	0,00	169,71
16	291,36	69,81	98,34	122,21	0,00	0,00	156,05
17	293,01	67,24	101,36	121,35	0,00	0,00	197,13
18	291,24	69,49	98,45	121,95	0,00	0,00	168,42
19	289,40	67,24	101,36	118,10	9,22	0,00	184,61
20	292,66	68,78	98,20	121,53	0,00	0,00	188,68
21	291,87	67,89	119,73	121,53	0,00	0,00	188,68
22	289,00	67,48	110,59	122,50	0,00	0,00	141,99
23	298,88	68,72	98,45	119,83	0,00	0,00	276,11
24	292,39	68,60	103,76	121,54	0,00	0,00	188,03
25	278,73	66,20	97,77	101,92	39,26	0,00	131,07
26	292,47	66,47	101,36	121,53	0,00	0,00	188,68
27	280,33	66,54	98,20	100,80	44,75	0,00	170,92
28	273,42	65,14	98,20	90,26	63,07	0,00	123,44
29	276,27	65,39	80,88	94,50	56,12	0,00	118,69

Table 5.3.2: Table including costs and design decision variables for the deterministic model with N representative days. Columns indicate: total cost for the tested solution on the entire period 2010-2020, total cost for the solution with N representative days, size of the photovoltaic system, size of the CHP system, size of the boiler, size of the electric energy system, size of the thermal energy system.

Differently from Section 5.2, here the minimum of the two curves coincides, for N=28, which is the best optimised solution. Note that, however, this does not imply it will be the best solution for the testing dataset too. Furthermore, this superimposition is a coincidence, which means solutions must always be tested, optimizing just the operation.

The trend of figure 5.3.6 is coherent, according to clustering theory, as for Section 5.2.

Finally, from figure 5.3.7 attention should go on the flatness of the curve up to N=25: one of the goal of data series reduction is to reduce computational effort, so a proper solution is to choose a lower number of representative days for the following comparisons. Repeating simulations many times, a good solution is N=11 days, even if the choice remains arbitrary.

5.3.3 Comparison between the two methods

An interesting point is to see which method between the one presented in Section 5.3.1 and the one in Section 5.3.2 is the best. In particular, the aim is to see which one is closer to the yearly reference solutions for the training dataset.

To do that, the first step is to evaluate the MAE and the RMSD for the optimized solutions obtained with clustering along all the training dataset. Clearly, for a given number N of days just one model is considered (clustering for the period 2005-2009), instead of five (yearly clustering), but the comparison is still done with respect to yearly reference models. Table 6.3.2 summarizes these results.

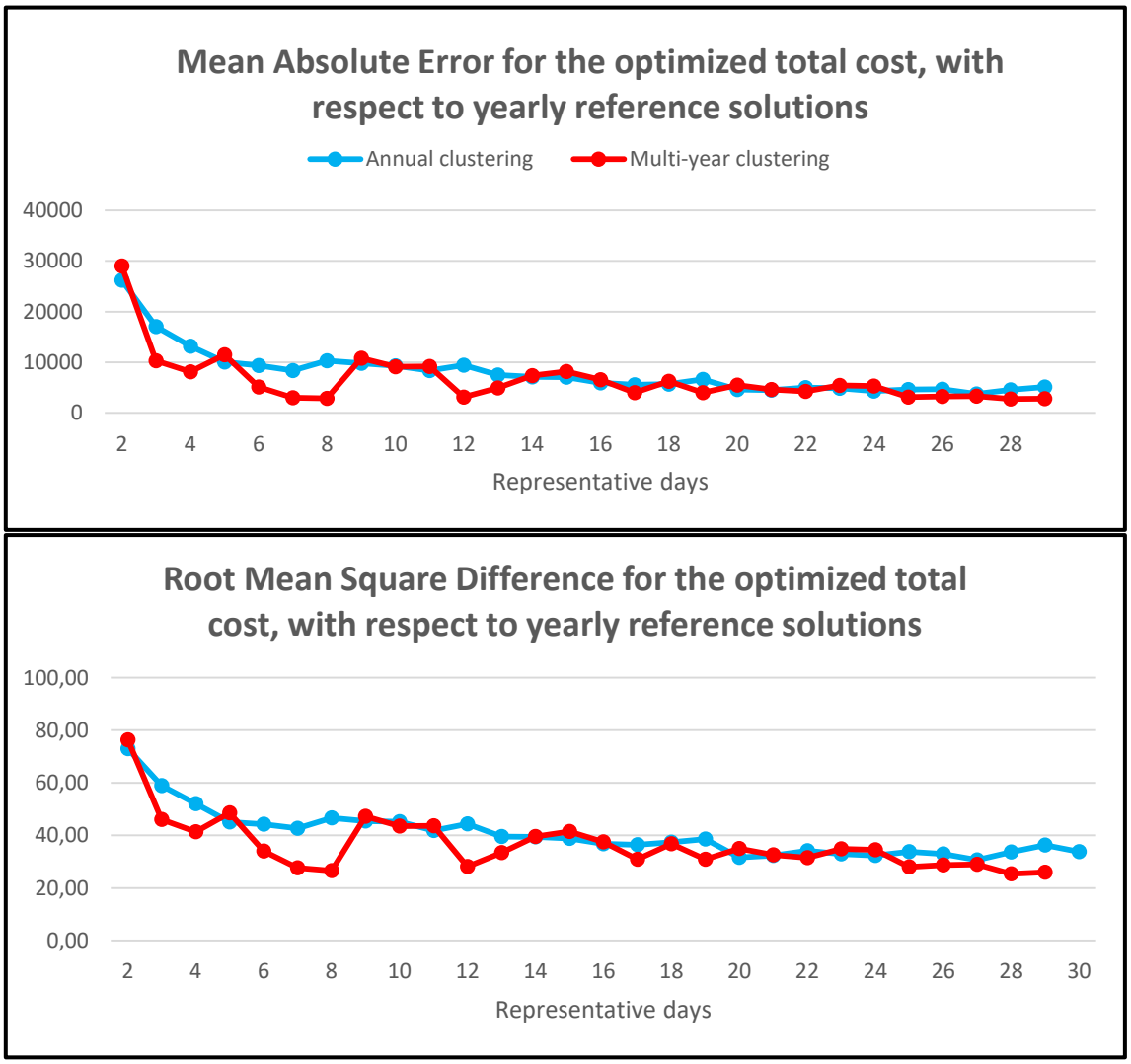

Figure 5.3.8: Comparison on Mean Absolute Error and Root Mean Square Difference for the deterministic models with N representative days, with respect to the yearly reference ones. In red the values from multi-year representative days, while blue indicates annual clustering.

As it is possible to see from figure 5.3.8, data series aggregation for the entire training dataset tends to minimize both MAE and RMSD, especially for a low number of representative days. In particular, this aspect is strongly relevant, because it allows decreasing sharply the computational weight of the problem and, consequently, the computational time, which is a great advantage.

The last advantage in using a single clustering process for the training dataset is given by exploitation of more information. In particular, if the adopted method is the yearly clustering one, once found the number of days

by minimizing MAE and RMSD, the actual days are taken from the closest year with respect to the yearly reference model. Therefore, there is a loss of information about days belonging to the other years of the training dataset. On the contrary, considering the complete training dataset for the clustering process, all data is used.

Please note that, as explained in previous Sections, the actual total cost can be found once tested the set of sizes obtained with the optimization process for the entire period. Here this comparison is not possible, because reference periods have different lengths and, consequently, one specific year could be particularly optimal for the PV production, which does not mean the solution obtained with the entire training dataset is worse, because it considers "worse" years too.

RMSD	MAE	0	1	2	3	4
		2005	2006	2007	2008	2009
		59848	61719	66644	62147	65892
76,37	29044	92294	92294	92294	92294	92294
46,06	10287	73537	73537	73537	73537	73537
41,31	8127	71377	71377	71377	71377	71377
48,53	11487	74737	74737	74737	74737	74737
33,96	5149	68399	68399	68399	68399	68399
27,61	3033	66042	66042	66042	66042	66042
26,51	2889	65621	65621	65621	65621	65621
47,21	10836	74086	74086	74086	74086	74086
43,56	9126	72376	72376	72376	72376	72376
43,67	9175	72425	72425	72425	72425	72425
28,13	3150	66237	66237	66237	66237	66237
33,46	4960	68210	68210	68210	68210	68210
39,56	7384	70634	70634	70634	70634	70634
41,43	8180	71430	71430	71430	71430	71430
37,56	6559	69809	69809	69809	69809	69809
30,86	3993	67242	67242	67242	67242	67242
36,77	6242	69492	69492	69492	69492	69492
30,84	3987	67237	67237	67237	67237	67237
34,95	5528	68778	68778	68778	68778	68778
32,60	4637	67886	67886	67886	67886	67886
31,51	4233	67483	67483	67483	67483	67483
34,81	5474	68724	68724	68724	68724	68724
34,49	5352	68602	68602	68602	68602	68602
28,02	3126	66197	66197	66197	66197	66197
28,75	3287	66466	66466	66466	66466	66466
28,96	3334	66543	66543	66543	66543	66543
25,33	2792	65138	65138	65138	65138	65138
25,93	2842	65388	65388	65388	65388	65388
		0	0	0	0	0

Year	Reference	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30

Table 5.3.3: MAE and RMSD, varying the number of representative days, for the multi-year clustering method. Columns indicate the objective function obtained as solutions of the deterministic models with N representative days, with respect to a specific year of the training dataset. The row "reference" indicates the reference model for the respective year. Columns MAE and RMSD indicate the Mean Absolute Error and the Root Mean Square Difference.

5.3.4 Optimization results with alternative ways to obtain representative days

The following results are referred to Section 4.3.3. Specifically, with "average" the method with average profiles of global solar irradiance and air temperature for each season (N=4) or month (N=12) is meant; with "season", the clustering technique to generate representative days is always the same, but the dataset is firstly divided per seasons. The comparison will be done with respect to results described in Section 5.3.1, focusing on the training dataset.

	C_TOT	*C_TOT_N*	*C_PV*	*C_CHP*	*C_boiler*	*E_EES*	*Q_TES*
	k€	**k€**	**kWp**	**kW**	**kW**	**kWh**	**kWh**
Seasonal N=4	296,66	73,57	67,59	122,50	0,00	45,88	141,97
Seasonal N=8	294,88	68,02	100,09	119,52	7,02	39,42	156,51
Seasonal N=12	291,29	70,00	123,70	121,70	0,00	0,00	180,28
Seasonal N=16	289,30	68,34	114,31	122,39	0,00	0,00	147,25
Seasonal N=20	281,60	68,08	101,31	107,99	35,26	0,00	132,91
Seasonal N=24	274,73	66,69	99,58	95,37	52,62	0,00	115,29
Seasonal N=28	276,19	65,90	103,95	95,67	50,39	0,00	139,24
Average seasonal	306,42	77,21	39,00	122,84	0,00	65,93	156,79
Average monthly	300,76	66,93	54,71	121,89	0,00	0,00	171,38

Table 5.3.4: Table including costs and design decision variables for the deterministic model with N representative days, generated with alternative methods. Columns indicate: total cost for the tested solution on the entire period 2010-2020, total cost for the solution with N representative days, size of the photovoltaic system, size of the CHP system, size of the boiler, size of the electric energy system, size of the thermal energy system.

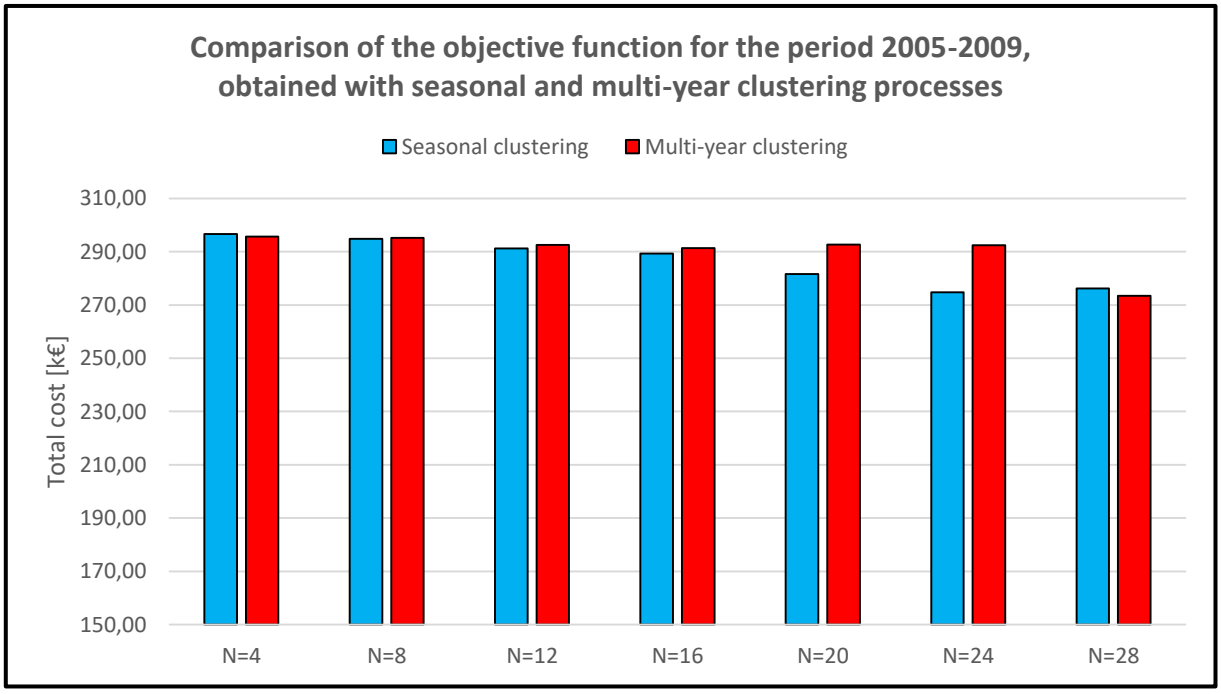

Figure 5.3.9: Comparison of the objective function for the period 2005-2009, obtained with seasonal and multi-year clustering processes. In particular, blue segments indicate use of clustering technique with previous division of the dataset in seasons, while the red ones do not divide it.

Looking at figure 5.3.9, the division of dataset in seasons before doing the clustering process allows obtaining lower costs for an "intermediate" number of representative days, included in the interval 12-24. In fact, for a low number of days, the necessity of having at least one day per season could imply a worse representation of the initial dataset, while for a higher number of days it reduces the overestimation (or underestimation) associated with clustering. Indeed, it forces to have the same number of days per season, which means there is an "equilibrium" in the consideration of other uncertain parameters, such as energy demands. For example, if N=12, 4 days are obtained both for summer and autumn; on the contrary, with the multi-year clustering process, more days represent summer than autumn. This aspect is good for the dataset representation, but it does not consider other input data, whose weight is crucial for the results.

However, if N increases too much, the possibility of avoiding rigid seasonal distinctions allows a reduction of the total cost, which makes the method of Section 5.3.2 to be the best.

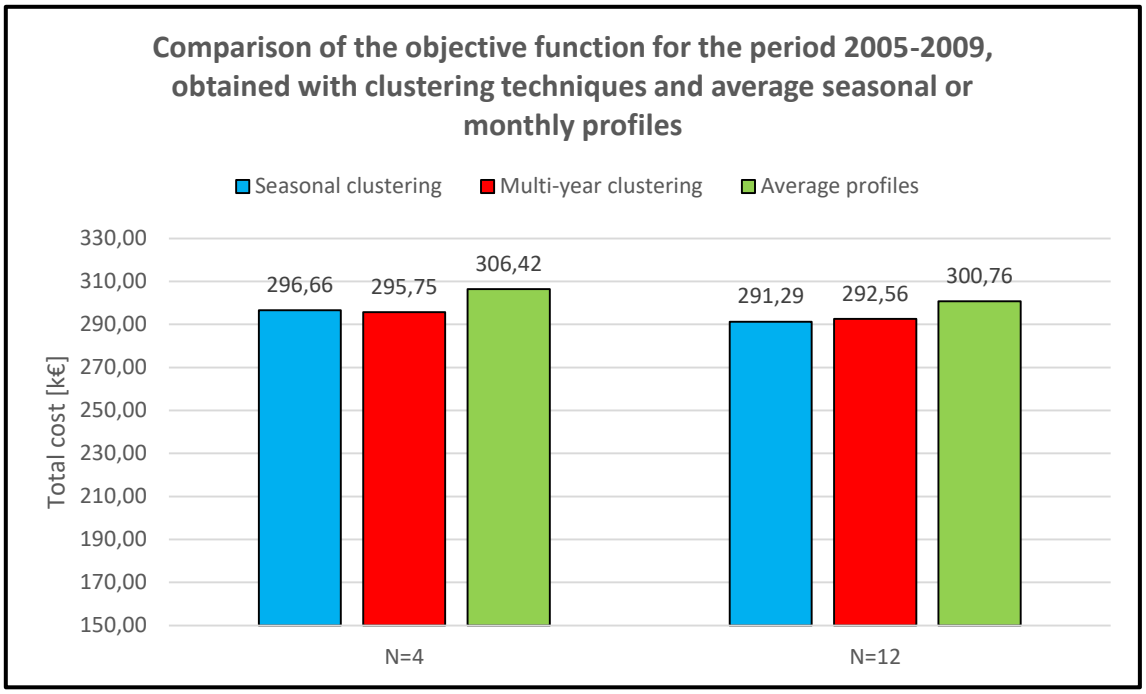

Figure 5.3.10: Comparison of the objective function for the period 2005-2009, obtained with clustering techniques and average seasonal profiles. Green segments indicate average profiles as input data: if N=4, these are seasonal profiles, while if N=12 they are monthly ones.

According to figure 5.3.10, the use of average profiles entails higher prices, compared to solutions with the same number of representative days. Nevertheless, remember that for N=11 or N=9, with multi-year clustering method, the solution would have a lower total cost. Many tries should be done with seasonal clustering method, to see if a different number of days per season could imply better solutions.

Let us look at table 5.3.4: if the number of days is lower than 20, the CHP size is always around 120kW, as already explained in Section 5.3.1, while it decreases if the boiler is introduced. This step, possible if N is high enough, implies a reduction in the total cost, because of the independence of each representative day. However, this means that the thermal side is more or less "constant" in terms of component sizes, so the difference is given mainly by the PV system.

If average profiles are adopted, there is an underestimation of the PV size, because the system overestimates the production along the day, avoiding peak conditions. In particular, if global solar irradiance is low, the production will be lower than the expected one, which means a higher size for PV would be better; if the global solar irradiance is high, a higher size (necessary to cover low irradiance conditions) would imply higher revenues.

5.4 Two-stage stochastic model of the residential MES with N representative days and M scenarios, for the period 2005-2009: objective function and decision variables

This Section puts attention on the results of the optimization process for the two-stage stochastic model developed in Section 4.3. In particular, the number of representative days is equal to 11, as explained in Section 5.3. The number of scenarios M, instead, varies from 2 to 7, to see which one is the best. Nevertheless, the limit is that each representative day has the same number of scenarios, independently on its weight, linked to the number of elements belonging to the cluster.

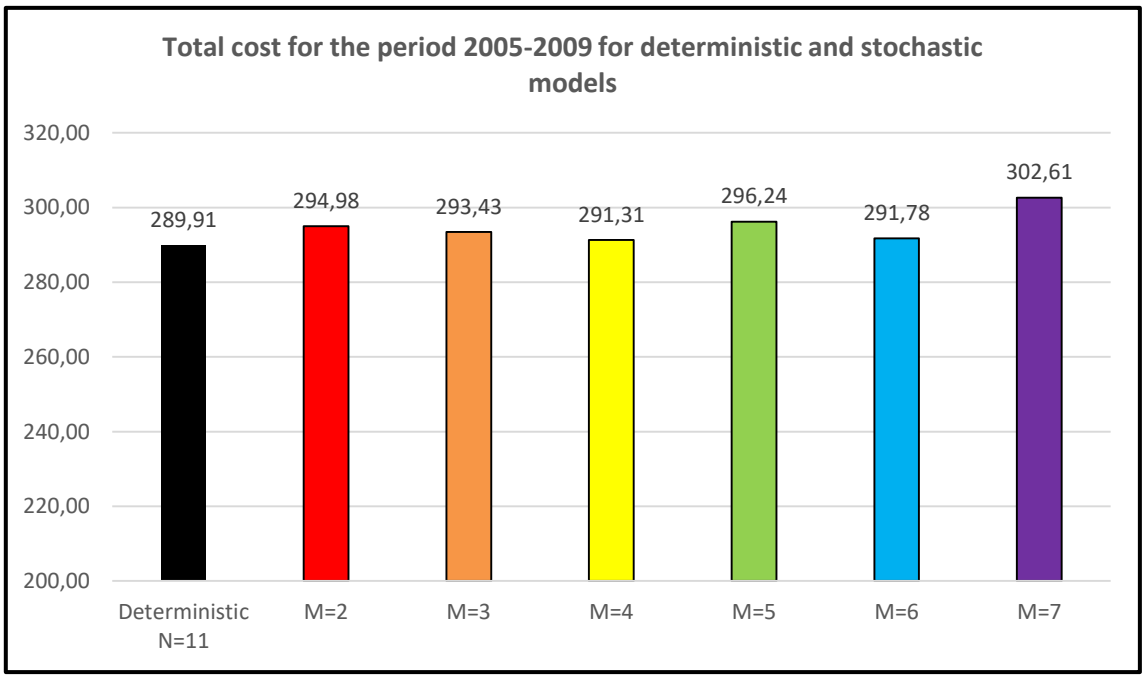

Figure 5.4.1: Total cost for the period 2005-2009 for deterministic and stochastic models. Each segment indicates the objective function tested on the entire period. The first one is obtained with the deterministic model with N representative days, while the other ones are from the two-stage stochastic model with M scenarios.

According to figure 5.4.1, even if the difference is not so high, the deterministic solution presents the lowest cost. Motivations can be found in table 5.3.4: the stochastic problem, including more scenarios for each representative day, tends to be give conservative solutions. For M=2, 3, 5 and 7, it underestimates the size of the PV system, preferring an EES to be more flexible if global solar irradiance is low. However, this entails higher costs. For M=4 or 6, on the contrary, it does not consider EES, but the overestimation of the TES implies a higher cost with respect to the deterministic solution. In any case, these two solutions give the lowest costs, because the ones with M=2, 3, 5 or 7 overestimate TES too. This could be related to the independence of each representative day and, furthermore, of each scenario. Nevertheless, conservative solutions are coherent, according to the literature about stochastic programming.

The entire set of sizes obtained by different number of scenarios will be used in the following Section, to see how good they are in the testing dataset.

	C_TOT	***C_TOT_N***	***C_PV***	***C_CHP***	***C_boiler***	***E _EES***	***Q _TES***
	k€	k€	kWp	kW	kW	kWh	kWh
Deterministic N=11	289,91	72,43	121,12	122,44	0,00	0,00	145,05
M=2	294,98	69,31	104,19	121,84	0,00	34,42	173,73
M=3	293,43	69,01	114,50	122,26	0,00	34,42	153,75
M=4	291,31	69,31	123,10	121,72	0,00	0,00	179,68
M=5	296,24	69,30	109,45	121,61	0,00	38,42	184,93
M=6	291,78	69,33	123,10	121,52	0,00	0,00	189,23
M=7	302,61	70,68	100,77	119,94	0,00	40,94	270,17

Table 5.3.4: Optimized costs and design decision variables for the two-stage stochastic model with N representative days and M scenarios. Columns indicate: total cost for the tested solution on the entire period 2010-2020, total cost for the solution with N representative days, size of the photovoltaic system, size of the CHP system, size of the boiler, size of the electric energy system, size of the thermal energy system. The first row is for the deterministic solution with 11 representative days, whereas the following ones for the stochastic solutions with M scenarios.

5.5 Comparison among all deterministic and stochastic models on the testing period 2010-2020

This Section compares results obtained with the deterministic and stochastic models described in Section 4, to see which one gives the closest solution to the reference one for the period 2010-2020, found in Section 5.2 with N=34 representative days.

Therefore, set of sizes obtained as optimization results for the different models are set, optimizing just the operation for the testing dataset, entailing 93690 time steps.

The first comparison is between the deterministic solution with 11 days as optimization period and the two-stage stochastic one with 11 days and M scenarios. According to figure 5.5.1, representing the relative error between the total cost of the given solution and the reference one, the deterministic solution presents the lowest relative error, equal to 5.86% (with N=9, the error would be 5,62%). Stochastic solutions, as for the training dataset, entail higher relative errors (higher total costs), for the motivation explained in Section 5.4. Among them, the lowest error is obtained for M=4, while the highest is for M=7, which is quite double the one obtained with the deterministic model.

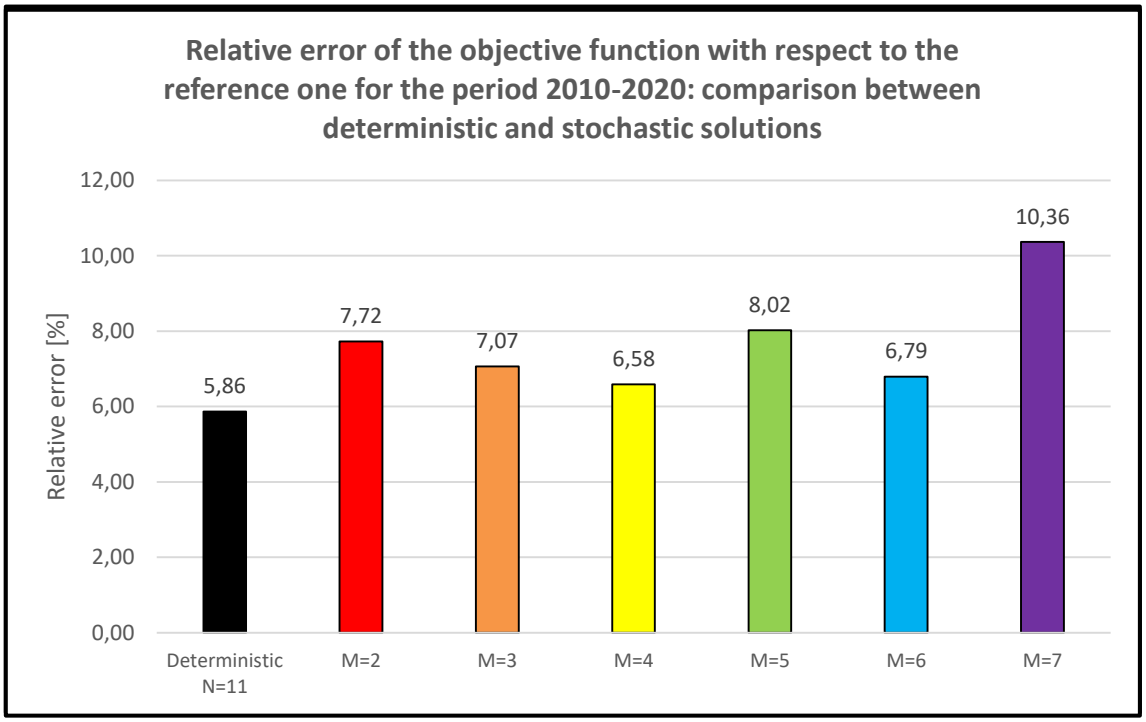

Figure 5.5.1: Relative error of the objective function with respect to the reference one for the period 2010-2020: comparison between deterministic and stochastic solutions. The first segment represents the deterministic solution with 11 representative days, while the other ones the stochastic solutions with M scenarios.

The second comparison considers the deterministic solutions obtained with N representative days, generated as explained in Section 4.3. In particular, “Multi-year clustering” indicates the use of clustering technique without considering seasonal division, for the entire dataset; “Average” indicates the use of monthly average profiles for the uncertain parameters; “Seasonal clustering” considers the use of clustering techniques with previous division per season of the dataset.

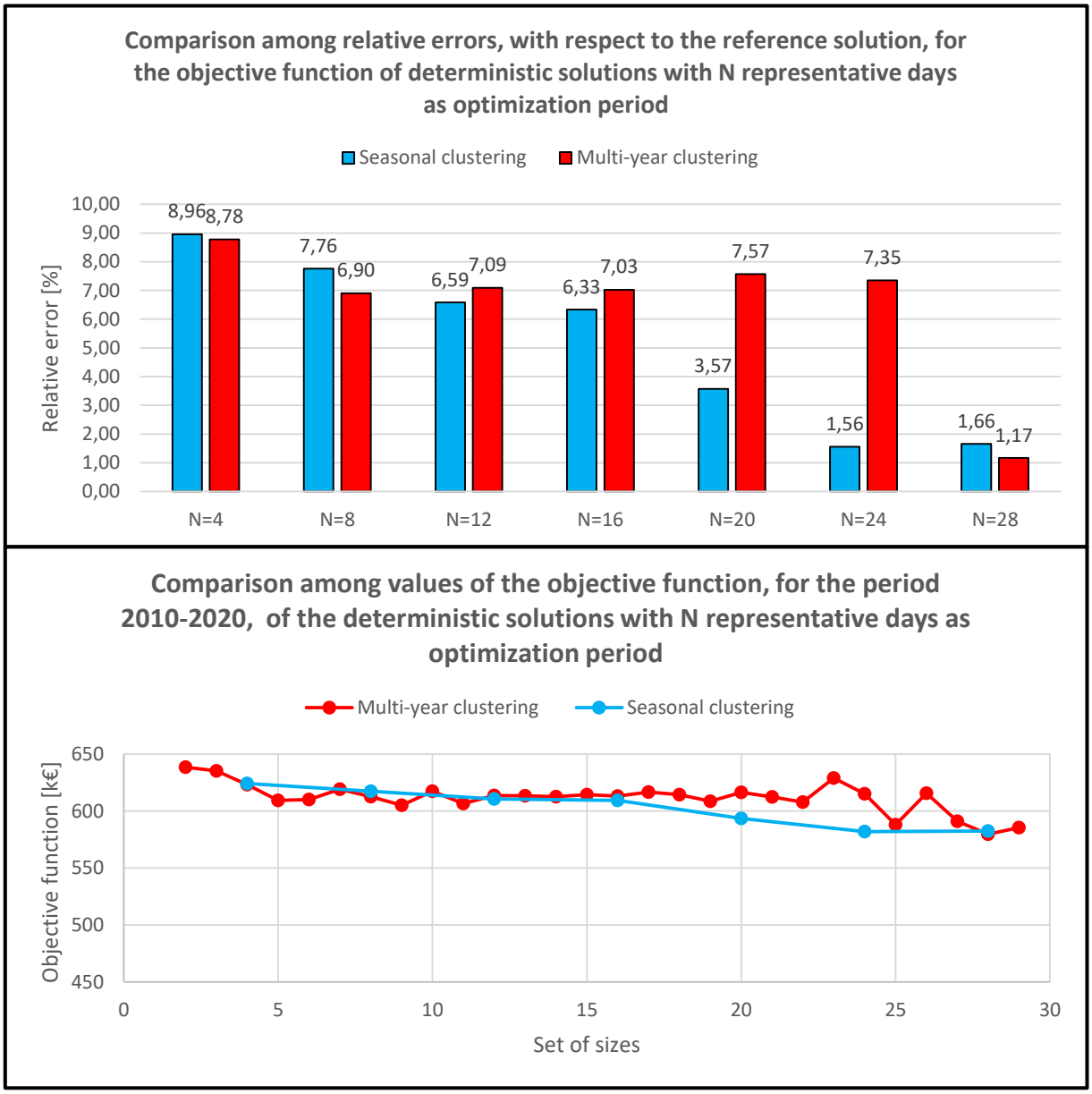

Figure 5.5.2: Relative error of the objective function with respect to the reference one for the period 2010-2020 (top) and values of objective function for the same period. This is a comparison between solutions obtained with seasonal and multi-year clustering techniques to generate representative days. In the second graph, the trend is indicated, even if values for the blue lines could oscillate: only cases with the same number of days per season are taken into account.

Results are the same as indicated in Section 5.3, according to figure 5.3.9: the seasonal clustering is more indicated for an “intermediate” number of representative days, because it forces having enough days per season. However, if N is low, equal to 4 or 8, or high, equal to 28, this aspect is a disadvantage. However, the necessity of introducing extreme thermal demand scenario entails similar solutions for N=4.

Moreover, let us remind that comparison, in the figure 5.5.2, are not in optimal conditions: for the multi-year clustering process, N=9 or 11, for instance, would be better than 8 or 12, as well as seasonal clustering could perform better with a different number of days per season.

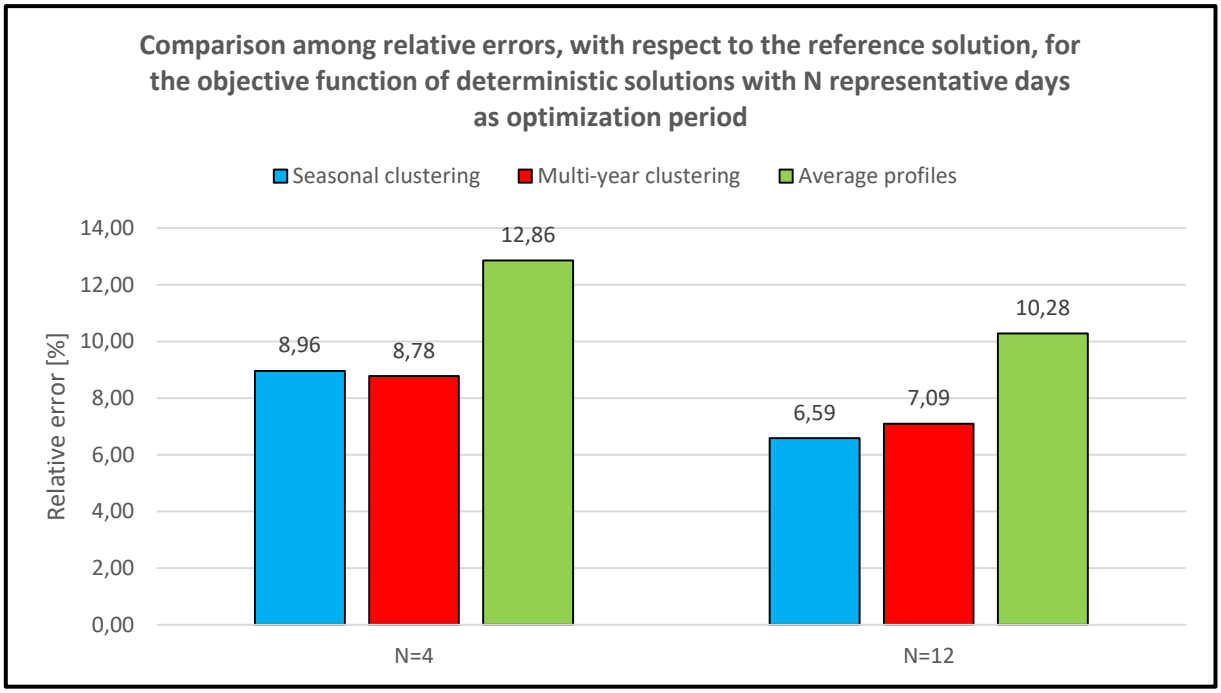

Figure 5.5.3: Comparison among relative errors, with respect to the reference solution, for the objective function of deterministic solutions with N representative days as optimization period. Blue segments represent seasonal clustering solutions, red ones multi-year clustering solutions and green ones solutions obtained with average seasonal or monthly profiles.

According to figure 5.5.3, the average profiles present the highest value of relative error, which means they are not suitable to be used. Moreover, let us remind that, for N=28, the other two methods involve a low error, lower than 2%. The difference in solutions could be even larger, but it is necessary to remind the introduction of the extreme scenario, which forces the thermal side to stay in certain ranges of sizes. Therefore, the huge difference is given by the PV system and the EES over the entire MES.

6. Conclusions

This thesis investigates the design and operation phases of an energy system under uncertainty. In particular, global solar irradiance and air temperature are the considered as uncertain parameters. The case study is of a residential multi-energy system, whose electrical and thermal requirement need to be satisfied. In particular, historical data for the period 2005-2009 (training dataset) are used to make design decisions and see their convenience for the period 2010-2020 (testing dataset).

Deterministic and two-stage stochastic models of such a system are developed to compare their solutions with the reference one for the period 2010-2020. Specifically, this reference solution is obtained under the assumption of perfect knowledge of the "future", which means the model should be solved with real data, for a total of 93690 time steps. However, due to computational complexity, clustering techniques are used. k-means is applied and the representative elements are those closest to the centroid, in order to avoid unreal average profiles, unless the worst case scenario of thermal demand belongs to the cluster.

The procedure is repeated with a variable number of representative days N, obtaining N-1 sets of sizes as results of the optimisation process. Then, the sizes are set and the operation is optimised for the whole period: the solution with the minimum total cost is the reference one. The reference value is thus obtained for N=34 representative days.

Emphasis is placed on the role of air temperature in the clustering process. A well-positioned element is a day whose season coincides with the season of the representative day of the cluster to which it belongs, which is crucial for its coupling with other input data, as thermal demand. If air temperature is included as a clustering attribute, the percentage of well-positioned elements increases by an average of 10%. Furthermore, comparing the same representative days for the case with only global solar irradiance and the one with it and air temperature, badly positioned elements are no longer present.

A deterministic model with N representative days as optimisation period is developed to compare its solutions with the reference ones. The days are generated starting from historical data for the period 2005-2009. Once the representative days are obtained, the optimisation process provides a set of sizes, which is tested for the training and testing datasets, optimising only the operation. Methods for generating representative days are:

- Clustering techniques to generate multi-year clusters. The procedure is the same as the one indicated for the reference model. One possibility is to generate annual clusters instead of multi-year ones, but this results in a loss of information and less accurate solutions, especially for a low number of representative days.
- Clustering techniques to generate seasonal clusters. The same clustering technique is used, but the dataset is first divided into seasons. Hence, the number of representative days is the same for each season. This method is found to be worse, with respect to the previous one, for a low number of representative days, equal to 4 or 8, or high, equal to 28. In particular, for N=4 or 8, the seasons are a constraint that could limit the representation of the dataset, while for N=28 they are not no longer interesting because they are a rigid distinction for the elements. For intermediate values of N, this method performs better because it reduces the overestimation (or underestimation) of sizes related to the imbalance in the dataset representation.
- Average seasonal (N=4) or monthly (N=12) profiles. It is shown that, optimising the operation, they present the worst relative error for the objective function, compared to the reference one. In particular, it is equal to 12,86% for the seasonal and 10,28% for the monthly, compared to 8,96% for the seasonal clustering and 8,78% for the multiannual, if N=4, or 6,59% and 7,09% if N=12. However, it should be noted that the best case for the first two methods is for N=28, with relative errors of 1,66% and 1,17% respectively.

Finally, an innovative two-step clustering process is adopted to generate scenarios for each representative day, developing a two-stage stochastic model with N days and M scenarios. First, representative days are generated using the same clustering technique, to get the same number of seasonal days as the deterministic model with 11 days, to obtain a fair comparison. Then, for each day a second clustering process is performed, aggregating elements belonging to such a cluster in order to generate different scenarios. However, once optimised the operation for the testing dataset, stochastic solutions prove to be the worst, with an average relative error of 7,76%, compared to the value of 5,86% of the deterministic solution. Stochastic solutions are more robust, according to stochastic theory, overestimating the TES and including the EES.

Future work could focus on the following points:

1. Study of the independence of the representative days. In fact, they are not temporarily linked, so the system is forced to let them independent on each other. However, it does not allow exploiting the link for the energy storages, implying oversized solutions.
2. Introduction of worst case scenarios. The adopted method entails oversizing of thermal side energy conversion units.
3. Introduction of prices of energy carriers and energy demand as uncertain parameters. Moreover, climate change or evolution of energy demand should also be included.

Bibliography

[1] Walker WE, Harremoës P, Rotmans J, Sluijs JP, van der, Asselt MBA, van, Janssen P, et al. Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support. Integrated Assessment 2003; Vol 4, No.1 pp.5–17. https://www.researchgate.net/publication/46641920_Defining_Uncertainty_A_Conceptual_Basis_for_Uncertainty_Management_in_Model-Based_Decision_Support

[2] AtomMirakyan, RolandDeGuio, Modelling and uncertainties in integrated energy planning Renewable and Sustainable Energy Reviews Volume 46, June 2015, Pages 62-69. https://www.sciencedirect.com/science/article/pii/S1364032115001161

[3] Cooke RM. Experts in uncertainty: opinion and subjective probability in science. USA: Oxford University Press; 1991.

[4] Georgios Mavromatidis, Kristina Orehounig, Jan Carmeliet, A review of uncertainty characterisation approaches for the optimal design of distributed energy systems. Renewable and Sustainable Energy revoews, Volume 88, May 2018, Pages 258-277. https://www.sciencedirect.com/science/article/pii/S1364032118300510

[5] Chong Zhang, Xue Xue, Qianzhou Du, Yimo Luo, Wenjie Gang, Study on the performance of distributed energy systems based on historical loads considering parameter uncertainties for decision making. Energy, Volume 176, 1 June 2019, Pages 778-791. https://www.sciencedirect.com/science/article/pii/S0360544219306668

[6] Georgios Mavromatidis, Kristina Orehounig, Jan Carmeliet, Uncertainty and global sensitivity analysis for the optimal design of distributed energy systems. Applied Energy, Volume 214, 15 March 2018, Pages 219-238. https://www.sciencedirect.com/science/article/pii/S0306261918300710

[7] R. Domínguez M. Carrión, G. Oggioni, Planning and operating a renewable-dominated European power system under uncertainty. Applied Energy, Volume 258, 15 January 2020, 113989. https://www.sciencedirect.com/science/article/pii/S0306261919316769

[8] Leonie Sara Plaga, Valentin Bertsch, Methods for assessing climate uncertainty in energy system models—A systematic literature review. Applied Energy, Volume 331, 1 February 2023, 120384. https://www.sciencedirect.com/science/article/pii/S0306261922016415

[9] Widén J, Carpman N, Castellucci V, Lingfors D, Olauson J, Remouit F, et al. Variability assessment and forecasting of renewables: a review for solar, wind, wave and tidal resources. Renewable and Sustainable Energy Reviews, Volume 44, April 2015, Pages 365-375. https://www.sciencedirect.com/science/article/pii/S1364032114010715

[10] Carta JA, Ramírez P, Velázquez S. A review of wind speed probability distributions used in wind energy analysis: case studies in the Canary Islands. Renewable and Sustainable Energy Reviews, Volume 13, Issue 5, June 2009, Pages 933-955. https://www.sciencedirect.com/science/article/pii/S1364032108000889

[11] Schnitzer M, Thuman C, Johnson P. Reducing uncertainty in solar energy estimates. AWSTruepower 2012. https://docplayer.net/15091232-Reducing-uncertainty-in-solar-energy-estimates.html

[12] Yingying Zheng, Bryan M. Jenkins, Kurt Kornbluth, Alissa Kendall, Chresten Træholt, Optimization of a biomass-integrated renewable energy microgrid with demand side management under uncertainty. Applied Energy, Volume 230, 15 November 2018, Pages 836-844. https://www.sciencedirect.com/science/article/pii/S0306261918313229

[13] Zahra Ghaemi, Thomas T. D. Tran, Amanda D. Smith, Sizing optimization of district energy systems considering meteorological, demand, and electricity emission factor uncertainties. https://engrxiv.org/preprint/view/1830/

[14] Jia Liu, Yuekuan Zhou, Hongxing Yang, Huijun Wu, Uncertainty energy planning of net-zero energy communities with peer-to-peer energy trading and green vehicle storage considering climate changes by 2050 with machine learning methods. Applied Energy, Volume 321, 1 September 2022, 119394. https://www.sciencedirect.com/science/article/pii/S0306261922007322

[15] Dongho Han, Jay H. Lee, Two-stage stochastic programming formulation for optimal design and operation of multi-microgrid system using data-based modeling of renewable energy sources. Applied Energy, Volume 291, 1 June 2021, 116830. https://www.sciencedirect.com/science/article/pii/S0306261921003299
[16] A.T.D. Perera, Vahid M. Nik, P.U. Wickramasinghe, Jean-Louis Scartezzini, Redefining energy system flexibility for distributed energy system design. Applied Energy, Volume 253, 1 November 2019, 113572. https://www.sciencedirect.com/science/article/pii/S0306261919312462
[17] Tayenne Dias de Lima, Alejandra Tabares, Nataly Banol Arias, John F. Franco, Investment & generation costs vs CO2 emissions in the distribution system expansion planning: A multi-objective stochastic programming approach. International Kournal of Electrical Power and Energy Systems, Volume 131, October 2021, 106925. https://www.sciencedirect.com/science/article/pii/S0142061521001654
[18] Zhuang Zheng, Xin Li, Jia Pan, Xiaowei Luo, A multi-year two-stage stochastic programming model for optimal design and operation of residential photovoltaic-battery systems. Energy and buildings, Volume 239, 15 May 2021, 110835. https://www.sciencedirect.com/science/article/pii/S0378778821001195
[19] G. Franke, M. Schneider, T. Weitzel, S. Rinderknecht, Stochastic Optimization Model for Energy Management of a hybrid microgrid using Mixed Integer Linear Programming. IFAC-PapersOnLine, Volume 53, Issue 2, 2020, Pages 12948-12955. https://www.sciencedirect.com/science/article/pii/S2405896320327841
[20]Ma, X.-Y., Sun, Y.-Z., Fang, H.-L., 2013. Scenario Generation of Wind Power Based on Statistical Uncertainty and Variability. IEEE Transactions on Sustainable Energy 4, 894–904. https://ieeexplore.ieee.org/document/6504817
[21]Georgios Mavromatidis, Kristina Orehounig, Jan Carmeliet, Design of distributed energy systems under uncertainty: A two-stage stochastic programming approach. Applied Energy, Volume 222, 15 July 2018, Pages 932-950. https://www.sciencedirect.com/science/article/pii/S0306261918305580
[22]Martin Wild, Doris Folini, Florian Henschel, Natalie Fischer, Bjorn Muller, Projections of long-term changes in solar radiation based on CMIP5 climate models and their influence on energy yields of photovoltaic systems. Solar Energy, Volume 116, June 2015, Pages 12-24. https://www.sciencedirect.com/science/article/pii/S0038092X15001668
[23] Prognos AG. Die Energieperspektiven fur die Schweiz bis 2050. Swiss Federal Office of Energy SFOE (Bundesamt für Energie BFE); 2012. https://www.prognos.com/sites/default/files/2021-01/120912_prognos_bundesamt_fuer_energie_energieperspektiven_schweiz_2050.pdf
[24] Meng Wang, Hang Yu, Xiaoyu Lin, Rui Jing, Fangjun He, Chaoen Li, Comparing stochastic programming with posteriori approach for multi-objective optimization of distributed energy systems under uncertainty. Energy, Volume 210, 1 November 2020, 118571. https://www.sciencedirect.com/science/article/pii/S0360544220316790
[25] IEA, World Energy Outlook 2022.
[26] IEA, Net Zero by 2050. A Roadmap for the Global Energy Sector.
[27] Holger Teichgraeber, Adam R. Brandt, Optimal design of an electricity-intensive industrial facility subject to electricity price uncertainty: Stochastic optimization and scenario reduction. Chemical Engineering Research and Design, Volume 163, November 2020, Pages 204-216. https://www.sciencedirect.com/science/article/pii/S026387622030441X
[28] Rolf Golombek, Arne Lind, Hans-Kristian Ringkjøb, Pernille Seljom, The role of transmission and energy storage in European decarbonization towards 2050. Energy, Volume 239, Part C, 15 January 2022, 122159. https://www.sciencedirect.com/science/article/pii/S0360544221024075
[29] Finn Roar Aune and Rolf Golombek, Are Carbon Prices Redundant in the 2030 EU Climate and Energy Policy Package? https://www.frisch.uio.no/publikasjoner/pdf/2021/Formatert/Aune_Golombek_1550_1.pdf
[30] Rafał Weron, Electricity price forecasting: A review of the state-of-the-art with a look into the future. International Journal of Forecasting, Volume 30, Issue 4, October-December 2014, Pages 1030-1081. https://www.sciencedirect.com/science/article/pii/S0169207014001083

[31] Iivo VehvilainenT, Tuomas Pyykkonen, Stochastic factor model for electricity spot price—the case of the Nordic market. Energy Economics, Volume 27, Issue 2, March 2005, Pages 351-367. https://www.sciencedirect.com/science/article/pii/S0140988305000034
[32] Paolo Gabrielli, Moritz Wüthrich, Steffen Blume, Giovanni Sansavini, Data-driven modeling for long-term electricity price forecasting. Energy, Volume 244, Part B, 1 April 2022, 123107. https://www.sciencedirect.com/science/article/pii/S036054422200010X
[33] Zheng Y, Jenkins BM, Kornbluth K, Træholt C. Optimization under uncertainty of a biomass-integrated renewable energy microgrid with energy storage. Renewable Energy, Volume 123, August 2018, Pages 204-217. https://www.sciencedirect.com/science/article/pii/S0960148118301307
[34] Yuehong Lu, Zhijia Huang, Ting Zhang, Method and case study of quantitative uncertainty analysis in building energy consumption inventories. Energy and Buildings, Volume 57, February 2013, Pages 193-198. https://www.sciencedirect.com/science/article/pii/S0378778812005634
[35] Deheng Zhou, Abu Bakkar Siddik, Lili Guo, Houjian Li, Dynamic relationship among climate policy uncertainty, oil price and renewable energy consumption—findings from TVP-SV-VAR approach. Renewable Energy, Volume 204, March 2023, Pages 722-732. https://www.sciencedirect.com/science/article/pii/S0960148123000241
[36] K. Gavriilidis, Measuring climate policy uncertainty[J], Available at SSRN 3847388 (2021). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3847388
[37] Ifedolapo Olabisi Olanipekun, Oktay Ozkan, Godwin Olasehinde-Williams, Is renewable energy use lowering resource-related uncertainties? Energy, Volume 271, 15 May 2023, 126949. https://www.sciencedirect.com/science/article/pii/S0360544223003432
[38] Holger Teichgraeber, Adam R. Brandt, Clustering methods to find representative periods for the optimization of energy systems: An initial framework and comparison. Applied Energy, Volume 239, 1 April 2019, Pages 1283-1293. https://www.sciencedirect.com/science/article/pii/S0306261919303022
[39] Leander Kotzur, Peter Markewitz, Martin Robinius, Detlef Stolten, Impact of different time series aggregation methods on optimal energy system design. Renewable Energy 117 (2018) 474e487. Renewable Energy, Volume 117, March 2018, Pages 474-487. https://www.sciencedirect.com/science/article/pii/S0960148117309783
[40]Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken wordrecognition. IEEE Trans Acoust Speech Signal Process 1978;26:43–9. https://doi. org/10.1109/TASSP.1978.1163055.
[41] Paparrizos J, Gravano L. k-shape: efficient and accurate clustering of time series. Acm Sigmod 2015:1855–70. https://doi.org/10.1145/2723372.2737793.
[42] G. Mavrotas, D. Diakoulaki, K. Florios, P. Georgiou, A mathematical programming framework for energy planning in services sector buildings under uncertainty in load demand: the case of a hospital in Athens, Energy Policy. ISSN: 03014215 36 (7) (2008) 2415e2429, https://doi.org/10.1016/j.enpol.2008.01.011.
[43] J.H. Merrick, On representation of temporal variability in electricity capacity planning models, Energy Econ. ISSN: 01409883 59 (2016) 261e274, https://doi.org/10.1016/j.eneco.2016.08.001.
[44] P. Nahmmacher, E. Schmid, L. Hirth, B. Knopf, Carpe diem: a novel approach to select representative days for long-term power system modeling, Energy. ISSN: 03605442 112 (2016) 430e442, https://doi.org/10.1016/j.energy.2016.06.081
[45] A.K. Jain, Data clustering: 50 years beyond K-means, Pattern Recognit. Lett. ISSN: 01678655 31 (8) (2010) 651e666, https://doi.org/10.1016/j.patrec.2009.09.011
[46] S.P. Adhau, R.M. Moharil, P.G. Adhau, K-Means clustering technique applied to availability of micro hydro power, Sustain. Energy Technol. Assess. ISSN: 22131388 8 (2014) 191e201, https://doi.org/10.1016/j.seta.2014.09.001.
[47] S. Fazlollahi, S.L. Bungener, P. Mandel, G. Becker, F. Mar_echal, Multi-objectives, multi-period optimization of district energy systems: I. Selection of typical operating periods, Comput. Chem. Eng. ISSN: 00981354 65 (2014) 54e66, https://doi.org/10.1016/j.compchemeng.2014.03.005.
[48] Stefan Pfenninger, Dealing with multiple decades of hourly wind and PV time series in energy models: A comparison of methods to reduce time resolution and the planning implications of inter-

annual variability, Applied Energy Volume 197, 1 July 2017, Pages 1-13, https://www.sciencedirect.com/science/article/pii/S0306261917302775
[49] Vinod HD. Integer programming and the theory of grouping. J Am Stat Assoc 1969;64(326):506–19. https://doi.org/10.1080/01621459.1969.10500990
[50] F. Domínguez-Munoz, J.M. Cejudo-Lopez, A. Carrillo-Andres, M. Gallardo- Salazar, Selection of typical demand days for CHP optimization, Energy Build. ISSN: 03787788 43 (11) (2011) 3036e3043, https://doi.org/10.1016/j.enbuild.2011.07.024.
[51] Peter J. Roussew, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, Volume 20, November 1987, Pages 53-65. https://www.sciencedirect.com/science/article/pii/0377042787901257
[52] Infanger, G., 1994. Planning Under Uncertainty: SolvingLarge-Scale Stochastic Linear Programs. Boyd & FraserPublishing Company, Stanford, CA.
[53] Gabriele Volpato, Gianluca Carraro, Luigi De Giovanni, Giovanni Andreatta, Andrea Lazzaretto, Enrico Dal Cin and Piero Danieli, A stochastic programming optimization framework to design an energy system and face market stages, https://www.research.unipd.it/handle/11577/3454157.
[54] I.L.R. Gomes, R. Melicio, V.M.F. Mendes, A novel microgrid support management system based on stochastic mixed-integer linear programming, Energy, Volume 223, 15 May 2021, 120030. https://www.sciencedirect.com/science/article/pii/S0360544221002796
[55] Gejirifu De, Zhongfu Tan, Menglu Li, Liling Huang, Xueying Song, Two-Stage Stochastic Optimization for the Strategic Bidding of a Generation Company Considering Wind Power Uncertainty
[56] Z. Zhou, J. Zhang, P. Liu, Z. Li, M.C. Georgiadis, E.N. Pistikopoulos, A two-stage stochastic programming model for the optimal design of distributed energy systems, Appl. Energy 103 (2013) 135–144, https://doi.org/10.1016/j.apenergy.2012.09.019.
[57] Chanwoong Jeon, Juneseuk Shin, Long-term renewable energy technology valuation using system dynamics and Monte Carlo simulation: Photovoltaic technology case. Energy, Volume 66, 1 March 2014, Pages 447-457. https://www.sciencedirect.com/science/article/pii/S0360544214000711
[58] Cong Chen, Xinwei Shen, Qinglai Guo, Hongbin Sun, Robust planning operation co-optimization of energy hub considering precise model of batteries' economic efficiency. Energy Procedia, Volume 158, February 2019, Pages 6496-6501. https://www.sciencedirect.com/science/article/pii/S1876610219301213
[59] Rong Li, Yong Yang, Multi-objective capacity optimization of a hybrid energy system in two-stage stochastic programming framework. Energy Reports, Volume 7, November 2021, Pages 1837-1846. https://www.sciencedirect.com/science/article/pii/S2352484721002079
[60] https://www.solsonica.com/ last access 21/04/2023
[61] http://www.enerblu-cogeneration.com/ last access 30/05/2023
[62] DEA-Danish Energy Agency. Technology Data. Available: https://ens.dk/en/our-services/projections-and-models/technology-data
[63] P. Sterchele, J. Brandes, J. Heilig, D. Wrede, C. Kost, T. Schlegl, et al., "Paths to a Climate-Neutral Energy System. The German Energy Transition in its Social Context," Fraunhofer ISE2020. https://www.ise.fraunhofer.de/en/publications/studies/paths-to-a-climate-neutral-energy-system.html
[64] https://www.arera.it/it/consumatori/consumatori_ele.htm last access 21/04/2023
[65] https://re.jrc.ec.europa.eu/pvg_tools/en/ last access 21/04/2023
[66] https://github.com/FZJ-IEK3-VSA/tsam, last access 18/08/2023
[67] http://scikit-learn.org/stable/modules/clustering.html, last access 18/08/2023
[68] Sergio Rech, Smart Energy Systems: Guidelines for Modelling and Optimizing a Fleet of Units of Different Configurations. *Energies* 2019, *12*(7), 1320; https://doi.org/10.3390/en12071320
[69] John A. Duffie et al. Solar Engineering of Thermal Processes, Photovoltaics and Wind, fifth edition
[70] F. Glover, "Improved Linear Integer Programming Formulations of Nonlinear Integer Problems," Management Science, vol. 22, pp. 455-460, 1975. http://www.jstor.org/stable/2630109

24.10.2023

Ringrazio la mia famiglia, che negli anni mi ha sempre supportato, senza mai dubitare di me.

Ringrazio Alessio, che mi ricorda spesso come un amico ti debba spingere a fare la cosa giusta, indipendentemente da quanto difficile essa sia. E indipendentemente da quanti litigi possa causare.

Ringrazio Daniele, che mi ha spronato a migliorarmi continuamente, sia come studente che come persona. Come studente, per non avermi mai dato tregua; come persona, per avermi insegnato ad essere paziente.

Ringrazio Riccardo, perché mi è stato vicino quando non mi doveva nulla, se non un paio di birre, e mi conosceva appena.

Ringrazio Michele, che per me è come un fratello, nonostante sia perennemente dall'altra parte del mondo.

Per ultima, ringrazio infinitamente Giulia, per essermi stata sempre accanto in questi otto anni, con pazienza e comprensione, ma soprattutto credendo ciecamente in me e obbligandomi ad uscire dalla mia pigrizia. Questo traguardo è tanto mio quanto tuo, perché sono la persona che sono grazie a te. Ti amo