Long-term scenario generation of heterogeneous random variables based on Markov chains for smart road tunnels planning

Long-term scenario generation of heterogeneous random variables based on Markov chains for smart road tunnels planning

Antonio Bracale
Department of Engineering
University of Naples Parthenope
Naples, Italy
antonio.bracale@uniparthenope.it

Pierluigi Caramia
Department of Engineering
University of Naples Parthenope
Naples, Italy
pierluigi.caramia@uniparthenope.it

Enrico Carpaneto
Department of Energy
Politecnico di Torino
Turin, Italy
enrico.carpaneto@polito.it

Pasquale De Falco
Department of Engineering
University of Naples Parthenope
Naples, Italy
pasquale.defalco@uniparthenope.it

Angela Russo
Department of Energy
Politecnico di Torino
Turin, Italy
angela.russo@polito.it

***Abstract*—This paper presents a methodology for the synthetic generation of heterogeneous random variables that affect the decision-making process in planning electric components and Distributed Energy Resources (DERs) for smart road tunnels. To ease the task of the decision maker, the methodology is developed in a generalized form in order to tackle variables of different sources and extractions (e.g., electrical, weather and economic variables). The methodology allows the generation of long-term synthetic scenarios of these heterogeneous variables through the development of Markov Chains (MCs), assuming that each variable can be represented by states with memory. The proposed methodology is applied and tested on actual data collected at the locations of two smart road tunnels in Italy, that are under evaluation to be equipped with Photovoltaic (PV) and wind generation systems and with a Battery Energy Storage System (BESS). Numerical experiments confirm the validity of the proposal for the synthetic generation of variables like solar irradiance, wind speed and electricity prices, as confirmed through the evaluation of their relevant statistics in a twenty-year period.**

Keywords—electricity price, Markov Chains, solar irradiance, synthetic scenario generation, wind speed

I. Introduction

Road tunnels are among the most energy-intensive mobility infrastructures. Strategically positioned along roadways, including in urban areas, they help reduce land consumption while enhancing the efficiency of the road network. The energy consumption of tunnels, driven by lighting systems, ventilation systems and auxiliary devices, results in significant costs and notable environmental impacts.

To enhance the sustainability of road tunnels, Distributed Energy Resources (DERs), like Photovoltaic (PV) and wind generation systems and Battery Energy Storage Systems, can be efficiently deployed. A key objective is to conceive the tunnel infrastructure integrating DERs as a microgrid [1] advancing toward the conceptual framework of Autonomous and Productive ROad Tunnels (APROT) [2].

The planning of APROTs plays a critical role in optimizing the sizing of installed DERs due to the significant capital investment required and the complex balance between economic, technical, and reliability factors, sometimes leading to reduced grid dependence. The planning process is inherently complex [3], as it must account not only for investment costs—primarily driven by the sizing of DERs—but also for operational expenses linked to their management, especially the charging and discharging strategies of BESSs. Additionally, intertemporal constraints, typically introduced by storage systems, must be carefully considered.

While traditional microgrid planning has often relied on deterministic models, numerous studies have shown that approaches capable of handling uncertainties are more suitable, given the increasing presence of uncertain parameters, variables, and data inputs. In the context of APROTs, sources of uncertainty may be associated with the tunnel electrical load, weather conditions affecting renewable energy production, economic factors, electricity prices, and other variables. Since the planning problem under consideration is long-term, uncertainty over the long term must be taken into account. In particular, in this paper, solar irradiance, wind speed and electricity prices are characterized over the planning horizon, assuming that the lighting and ventilation system loads are not affected by uncertainty against time.

Generating synthetic long-term scenarios is a viable approach to understand and mitigate the problems caused by the uncertainty of the random variables affecting APROT planning. Basing the decision-making process on artificial data, rather than using actual measurements, offers several advantages, including flexibility (the scenario generation avoids biases toward specific past situations that are not certain to occur again in the future) and generalization (the scenario generation has no limit in terms of data, nor in terms of the length of the period included in the decision-making analysis) [4].

Various scenario generation methods have been proposed in the relevant literature, each one based on different approaches and with different features. Following the classification presented in [5], the methods can be broadly classified into sampling-based methods, forecasting-based methods, and optimization-based methods. Sampling-based

This study was carried out within the project "Models and methods for SUstainable Energy Road Tunnels in microgrid Environment able TO provide Grid service and External ENergy (SUERTE TO GREEN)" funded by EU-NextGenerationEU plan through the MUR "Bando Prin 2022-DD 1409 14/09/2022". This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

methods create discrete scenarios by drawing samples from the probability distributions of the target variable and utilizing the resulting outputs as generated scenarios. Examples of such methods include Monte Carlo simulations, Latin Hypercube Sampling, copula function sampling, Markov Chains, sample average approximation, and persistence and variation techniques. Forecasting-based methods, on the other hand, rely on training models with historical data to generate scenarios without making assumptions about the underlying distribution functions. Notable examples include Seasonal AutoRegressive Integrated Moving Average models, state-space methods, and (deep) neural networks. Optimization-based methods differ from the previous two categories by focusing on reducing the number of scenarios from a large set of possibilities. These methods typically aim to minimize the distance between the selected subset of scenarios and the full set of potential scenarios. Depending on the approximation criteria, optimization-based approaches include clustering techniques, backward reduction methods, forward selection strategies and moment-matching techniques.

Among the proposed methods, Markov Chains (MCs), employed within a probabilistic framework to analyse transitions between states over time, have been demonstrated to be particularly well-suited for long-term characterization as they are able to capture temporal dependencies and the stochastic nature of uncertain variables while simplifying complexity; in fact, future states of the MC depend only on a limited number of preceding states rather than the entire historical record. Studies that utilize MC-based methodologies to generate synthetic wind speed samples are [6-10]. Examples of methodologies utilizing MCs to generate synthetic solar irradiance samples can be found in [4,11-13]. To the best of our knowledge, MC-based approaches have been applied for generating long-term electricity price series only in [14].

This paper extends the use of MCs to generate long-term synthetic scenarios of heterogeneous random variables that may be involved in the decision-making process for APROT planning. The methodology proposed in this paper is based on frequentist MCs, in which Higher-Order MCs (HOMCs) are developed with transition probabilities estimated under diversified approaches (namely the traditionally used Full Frequentist Approach (FFA), traditionally used in MC-based methods and presented in [14], and the novel Segmented Frequentist Approach (SFA)). The SFA is specifically introduced in this paper to avoid estimating the transition probabilities on too few occurrences, which is a problem in HOMCs with large orders that may lead to a too sparse discretization of data into the bins (i.e., the intervals corresponding to the MC states). The proposed methodology aims to be unified and applied for the generation of synthetic scenarios of the different and heterogeneous random variables that are involved in the decision-making process for APROT. It is specifically formulated to suit heterogeneous random variables of different nature and extraction, and two identification strategies of states are considered (namely the Uniform Binning (UB) strategy, traditionally used in MC-based methods such as the one in [14], and the Non-Uniform Binning strategy (NUB)), to also model data that frequently assume a specific value in correspondence to their minimum value (e.g., the solar irradiance, that is typically used for assessing the PV producibility, is frequently zero).

The main contributions of this paper can be summarized as: i) the introduction of the SFA to solve the HOMC transition probabilities estimation; ii) the introduction of the NUB strategies to enhance the modeling of heterogeneous random variables. These contributions are specifically tuned to the application of the methodology to variables involved in the APROT planning decision-making processes and to further extend the use of MC-based methodology to generate long-term scenarios for electricity prices, that is only at an early stage of research.

This paper is organized as it follows. Section II recalls the APROT structure, Section III presents the proposed MC-based methodology, and Section IV illustrates the numerical applications. Conclusions are in Section V.

II. Background: Autonomous and Productive Road Tunnels (APROTs)

The APROT is conceived to operate as a Microgrid (MG), running most of the time in a grid-connected mode but occasionally performing in islanded mode during particular contingencies.

The sample layout of the considered APROT system is illustrated in Fig. 1. The loads of the tunnel (mostly made of ventilation and lighting systems) are supplied by the upstream network and/or by the local distributed generation systems. A BESS may operate either to guarantee reliability/security levels (e.g., by maintaining a sufficient state of charge to supply the tunnel loads for some minutes during contingencies), to minimize the operating costs, or to maximize the exploitation of the energy produced locally.

We highlight that the BESS management strategy has a strong impact on the formulation of the optimal planning problem to adequately size the components of the APROT, but this is beyond the scope of this paper. We also consider that the tunnel loads mostly consist of lighting and ventilation systems, thus they can be assumed to be almost invariant in terms of uncertainty throughout the planning period.

Under these assumptions, the problem addressed here is that the decision maker in charge of the optimal planning of these devices should get information about DER capabilities (mostly linked by the availability of the weather-driven primary energy source) and electricity prices on the market through the expected lifetime of the investment (conventionally set here at 20 years). These heterogeneous variables are intrinsically uncertain and cannot be reliably predicted through such a long horizon.

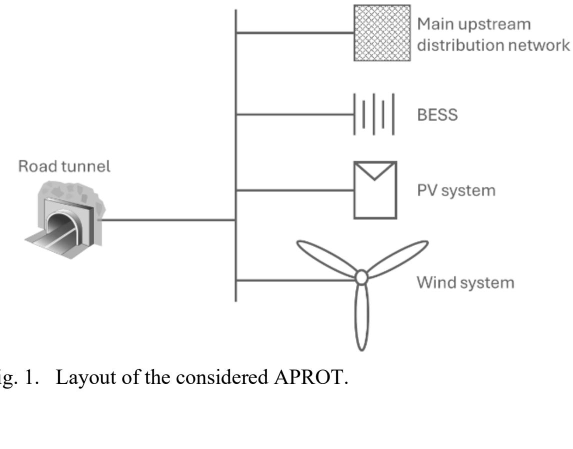

Fig. 1. Layout of the considered APROT.

The problem can however be solved through the generation of synthetic scenarios of these random variables affecting the APROT optimal planning, in order to enhance the flexibility (the scenario generation avoids biases toward specific past situations that are not certain to occur again in the future) and the generalization (the scenario generation has no limit in terms of data, nor in terms of the length of the period) of the decision-making process. This is achieved in this paper through the proposed frequentist MC-based methodology, described in the next Section.

III. The Proposed Markov Chain-Based Methodology for the Synthetic Long-Term Scenario Generation

The methodology proposed in this paper for the synthetic generation of scenarios of heterogeneous random variables is based on frequentist MCs. As it is well known, a MC assumes that the random variable can be characterized through the probability of belonging to states, i.e., of assuming values within specified bins.

Below we formulate the methodology on the generic uncertain variable x that is assumed to be characterized through N states $S^{\langle 1\rangle}, S^{\langle 2\rangle}, \dots, S^{\langle N\rangle}$, where the generic j^{th} state $S^{\langle j\rangle}$ corresponds to a bin $]l^{\langle j\rangle}, u^{\langle j\rangle}]$ and where $l^{\langle j\rangle}, u^{\langle j\rangle}$ are the bin's lower and upper bound, respectively. Note that, although $l^{\langle j\rangle} \neq u^{\langle j\rangle}$ in most of the existing applications, here we do not apply this constraint, allowing the creation of "collapsed" bins that include only one value (i.e., singletons). This is particularly important to generalize the methodology to random variables that frequently assume a specific value, as in the case of solar irradiance (which is zero in many periods of the day).

Anyway, the criterium for staying in a state is based on the value x_t assumed by the variable x at a generic time t: if $l^{\langle j\rangle} < x_t \leq u^{\langle j\rangle}$, the variable is in the j^{th} state $S^{\langle j\rangle}$ at time t. This is represented in this paper as $S_t = S^{\langle j\rangle}$.

A. Recalls on First-Order Markov Chains (FOMCs) and Higher-Order Markov Chains (HOMCs)

MCs characterize the underlying stochastic processes that determine the time-varying nature of a random variable, considering that the probability of the variable to be in a given state in the future depends on the occurrence of past states [15,16]. This "memory" is however limited to the most recent states, and it is linked to the MC order. In a HOMC with order o, the property of independence from all but the last o states determines that the state S_t at time t depends only on the previous o states $S_{t-k}, S_{t-2k}, \dots, S_{t-ok}$:

$$p(S_t|S_{t-k}, S_{t-2k}, \dots, S_1) = p(S_t|S_{t-k}, S_{t-2k}, \dots, S_{t-ok}) \tag{1}$$

where k is the step of the time series modeled by the MC, and where the process is defined by the transition probabilities among the states. Below we respectively represent as $a_{i_o i_{o-1} \dots i_2 i_1 j}$ and as $\hat{a}_{i_o i_{o-1} \dots i_2 i_1 j}$ the true and the estimated probability of the variable to transition among the states $S^{\langle i_o\rangle}, S^{\langle i_{o-1}\rangle}, \dots, S^{\langle i_2\rangle}, S^{\langle i_1\rangle}, S^{\langle j\rangle}$, at times $t-ok$, $t-(o-1)k, \dots, t-2k,\ t-k,\ t$, respectively.

The true probabilities are often unknown as the underlying stochastic process is unknown, but the set of N^o estimated probabilities, which is a $N \times N \times \dots \times N$ tensor $\widehat{\boldsymbol{A}}$, can be used to predict the probability of the state in which the variable will stay in a future time t. For example, the predicted probability $\hat{\pi}_t^{\langle j\rangle}$ of the variable staying in the j^{th} state at time t is obtained as [16]:

$$\hat{\pi}_t^{\langle j\rangle} = \sum_{i_o=1}^{N} \sum_{i_{o-1}=1}^{N} \dots \sum_{i_2=1}^{N} \sum_{i_1=1}^{N} (\hat{a}_{i_o i_{o-1} \dots i_2, i_1 j} \cdot \pi_{t-ok}^{\langle i_o\rangle} \cdot \pi_{t-(o-1)k}^{\langle i_{o-1}\rangle} \cdot \dots \cdot \pi_{t-k}^{\langle i_1\rangle}) \tag{2}$$

where the generic probability $\pi_t^{\langle i\rangle}$ corresponds to the event in which the variable at time t stays in the i^{th} state. The probability values appearing in (2) are therefore either one or zero, respectively when the observed value of the variable falls or does not fall within the i^{th} state. For example, if the value of the variable x_{t-4} satisfies the condition $l^{\langle 8\rangle} < x_{t-4} \leq u^{\langle 8\rangle}$, the variable is in the eighth state (i.e., $S_{t-4} = S^{\langle 8\rangle}$), and thus $\pi_{t-4}^{\langle 8\rangle} = 1$ and $\pi_{t-4}^{\langle i\rangle} = 0$ for any $i \neq 8$.

With this position, the prediction statement (2) is fully dependent on how the transition probabilities $\hat{a}_{i_o i_{o-1} \dots i_2 i_1 j}$ are estimated. This problem is addressed in the following sub-Sections that particularize the proposed model to make a general approach that can apply to heterogeneous variables, either with a FFA or with a SFA.

We eventually highlight that the above formulation collapses into the First-Order Markov Chains (FOMCs) when $o = 1$. In such case, (1) and (2) become:

$$p(S_t|S_{t-k}, S_{t-2k}, \dots, S_1) = p(S_t|S_{t-k}) \tag{3}$$

$$\hat{\pi}_t^{\langle j\rangle} = \sum_{i=1}^{N} \hat{a}_{ij} \cdot \pi_{t-k}^{\langle i\rangle} \tag{4}$$

Below we present the methodology with general reference to HOMCs, but it can be trivially particularized to the simpler case of FOMCs.

B. Full Frequentist Approach (FFA)

In the FFA, the transition probabilities are estimated by considering the number of observed transitions among the states in a training window. In particular, if the variable transitioned $\widehat{\#}_{i_o i_{o-1} \dots i_2 i_1 j}$ times among the states $S^{\langle i_o\rangle}, S^{\langle i_{o-1}\rangle}, \dots, S^{\langle i_2\rangle}, S^{\langle i_1\rangle}, S^{\langle j\rangle}$ (in this exact order), and it transitioned $\widehat{\#}_{i_o i_{o-1} \dots i_2, i_1}$ times among the states $S^{\langle i_o\rangle}, S^{\langle i_{o-1}\rangle}, \dots, S^{\langle i_2\rangle}, S^{\langle i_1\rangle}$ (in this exact order), then the transition probability $\hat{a}_{i_o i_{o-1} \dots i_2 i_1 j}$ is estimated as:

$$\hat{a}_{i_o i_{o-1} \dots i_2 i_1 j} = \begin{cases} \frac{\#_{i_o i_{o-1} \dots i_2 i_1 j}}{\#_{i_o i_{o-1} \dots i_2 i_1}} & \text{if } \#_{i_o i_{o-1} \dots i_2 i_1} \neq 0 \\ \frac{1}{N} & \text{if } \#_{i_o i_{o-1} \dots i_2 i_1} = 0 \end{cases} \tag{5}$$

Note that, with such estimation, $\sum_{j=1}^{N} \hat{a}_{i_o i_{o-1} \dots i_2 i_1 j} = 1$ for any previous combination of states $S^{\langle i_o\rangle}, S^{\langle i_{o-1}\rangle}, \dots, S^{\langle i_2\rangle}, S^{\langle i_1\rangle}$, thus fully covering the probability space. Note also that the second statement in (5) corresponds to the uninformed, uniform selection of probabilities when no recorded transition has ever occurred among states $S^{\langle i_o\rangle}, S^{\langle i_{o-1}\rangle}, \dots, S^{\langle i_2\rangle}, S^{\langle i_1\rangle}$.

C. Segmented Frequentist Approach (SFA)

In the SFA, the transition probabilities are estimated by considering only the number of observed transitions among the first and the last states, in a training window. In particular, if the variable transitioned $\widehat{\#}_{i_o j}$ times from the state $S^{\langle i_o\rangle}$ towards the state $S^{\langle j\rangle}$ after o steps, whatever the intermediate states were, and it was observed $\hat{\#}_{i_o}$ times in state $S^{\langle i_o \rangle}$, then the transition probability $\hat{a}_{i_o i_{o-1} \dots i_2 i_1 j}$ is estimated as:

$$\hat{a}_{i_o i_{o-1} \dots i_2 i_1 j} = \begin{cases} \frac{\hat{\#}_{i_o j}}{\hat{\#}_{i_o}} & \text{if } \#_{i_o} \neq 0 \\ \frac{1}{N} & \text{if } \#_{i_o} = 0 \end{cases} \quad (6)$$

for any intermediate states $i_{o-1}, \dots, i_2$. This is conceptually equivalent to neglect the impact of the intermediate states in the estimation of the transition probabilities, maintaining only the first and the last observed states as relevant.

D. *Identification of the States for Continuous Random Variables*

We hereby assume that the heterogeneous variables to be synthetically generated are continuous random variables, thus it is necessary to identify and optimize the states in which the continuous domain is discretized. In this paper, we apply two alternative strategies for the identification of states, namely the UB and the NUB, which is developed with a Lower Singleton (NUB-L). Assuming that N is the number of states:

i) in the UB the generic j^{th} bin's lower and upper bound $l^{\langle j \rangle}, u^{\langle j \rangle}$ are set for $j = 1, \dots, N$ as:

$$l^{\langle j \rangle} = x_{min} + \frac{(j-1)}{N}(x_{max} - x_{min}) \quad (7)$$

$$u^{\langle j \rangle} = x_{min} + \frac{j}{N}(x_{max} - x_{min}) \quad (8)$$

where $x_{min} = \min\{x_t, t \in \Omega_{tra}\}$, $x_{max} = \max\{x_t, t \in \Omega_{tra}\}$, and Ω_{tra} is the set of time indices corresponding to the training window. Note that this position corresponds to a discretization in which each of the N bins has the same length $(x_{max} - x_{min})/N$, and they span the whole continuous domain of the target variable observed during the training window;

ii) in the NUB-L the generic j^{th} bin's lower and upper bound $l^{\langle j \rangle}, u^{\langle j \rangle}$ are set for $j = 2, \dots, N$ as:

$$l^{\langle j \rangle} = x_{min} + \frac{(j-2)}{N-1}(x_{max} - x_{min}) \quad (9)$$

$$u^{\langle j \rangle} = x_{min} + \frac{(j-1)}{N-1}(x_{max} - x_{min}) \quad (10)$$

while the first bin is a singleton with $l^{\langle 1 \rangle} = u^{\langle 1 \rangle} = x_{min}$.

Note that, in the NUB-L strategy, all the bins except the singletons have the same length, i.e., $(x_{max} - x_{min})/(N-1)$. However, the NUB-L strategy gives the possibility to represent random variables that frequently assume a specific value only in correspondence to their minimum value.

E. *Optimal Selection of the MC Order and Number of States and Long-Term Synthetic Scenario Generation*

We perform a grid search to optimize the MC order and number of states, checking the performance of different MCs with different architectures in a validation window. This grid search iterates through values $1 \leq o \leq o^*$, i.e., up to a preassigned maximum order o^*, and through values $1 \leq N \leq N^*$, i.e., up to a preassigned maximum number of states N^*.

In the training stage, for an assigned couple (o, N), the tensor $\widehat{\boldsymbol{A}}(o, N)$ is estimated either using the FFA or the SFA and short-term predictions $\hat{x}_t(o, N)$ are issued by (2) throughout the validation period (i.e., for $t \in \Omega_{val}$ where Ω_{val} is the set of time indices corresponding to the validation window). The best MC architecture, in terms of optimal $\bar{o}, \bar{N}$ values, is selected by minimizing the Mean Squared Error (MSE) in the validation period as:

$$\bar{o}, \bar{N} = \underset{o,N}{\operatorname{argmin}} \frac{1}{T_{val}} \sum_{t \in \Omega_{val}} [\hat{x}_t(o, N) - x_t]^2 \quad (11)$$

Long-term synthetic scenarios are eventually generated through an iterative procedure employing the optimized MC. For each scenario, the MC model is propagated step-by-step across the entire planning duration (hourly for Y years, with time steps $t \in \Omega_{lts}$ where Ω_{lts} is the set of time indices corresponding to the long-term synthetic generation). For the generic l^{th} scenario, at each time step t, the following sequence is executed: i) next-step probabilities $\hat{\pi}_t^{\langle 1 \rangle}, \hat{\pi}_t^{\langle 2 \rangle}, \dots, \hat{\pi}_t^{\langle \bar{N} \rangle}$ are predicted using (2); ii) a random number is drawn from a uniform distribution and a predicted state $\hat{S}_t$ is determined by applying inverse sampling to the cumulative predicted probabilities; iii) future state probabilities are assigned according to the predicted state $\hat{S}_t$ (e.g., if $\hat{S}_t = S^{\langle j \rangle}$, $\pi_t^{\langle j \rangle} = 1$ and $\pi_t^{\langle i \rangle} = 0$ for any $i \neq j$), ensuring the MC can proceed to the next time step. Note that the diversity in the l^* generated scenarios comes from the different draws from the uniform distribution, that lead to different predicted states.

Fig. 2 shows the diagram flow of the MC architecture optimization and of the procedure for the generation of the synthetic scenarios.

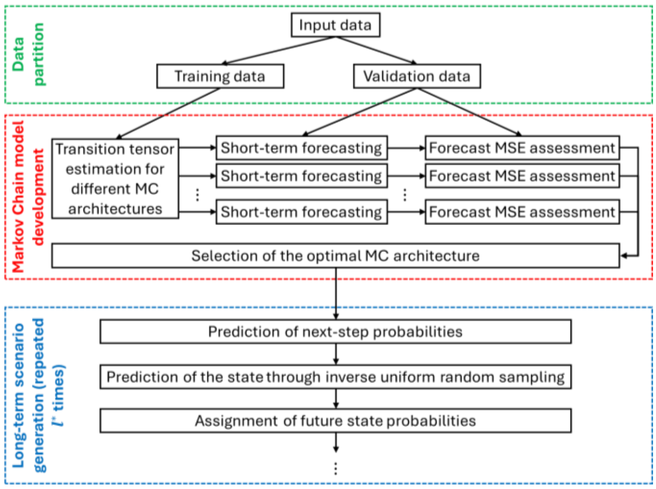

Fig. 2. Diagram flow of the MC architecture optimization and of the procedure for the generation of synthetic scenarios.

IV. Numerical Applications

The proposed methodology is applied to generate long-term synthetic scenarios of three continuous but heterogeneous variables that are involved in the APROT planning and decision-making process: the solar irradiance (to assess the producibility of PV systems), the wind speed (to assess the producibility of wind turbines), and the electricity prices (to quantify the value of the energy traded between the APROT and the upstream distribution network). These variables are specifically selected in this paper both since they are very impactful on the planning performance and since they have different characteristics, which is important for evaluating our methodology. In particular, solar irradiance values have a strong daily and yearly seasonality and they are frequently equal to the lower bound of their domain (i.e., zero, exemplarily during nighttime hours), and this can help in understanding the role of the NUB-L in the development of the MCs. Wind speed values are weather-driven, too, but they only typically have a moderate/mild yearly seasonality. Electricity prices are instead driven by multiple factors (e.g., weather, economic, and social ones) but they often show a strong daily, weekly and yearly seasonality.

Actual data of these three variables are used in the numerical applications presented in this paper. They are collected at an Italian site located in the "South" electricity market bidding area. Weather data are retrieved from the ERA5 reanalysis database [17] and prices are retrieved from the ENTSO-E database [18], at hourly resolution through the years 2016-2019 (35064 samples for each variable). The training window covers the first three observed years (i.e., $\Omega_{tra} = \{1, \dots, 26304\}$), the validation window covers the last observed year (i.e., $\Omega_{val} = \{26305, \dots, 35064\}$), and the test window spans the next twenty years, to replicate the typical period considered in investments for power system planning.

The number of generated synthetic scenarios is $l^* = 2000$ for all variables. The methodology is independently applied to each hour of the day, in order to enhance the performance of the MCs particularly for variables showing a strong daily seasonality. This means that, exemplarily, a dedicated MC is developed and optimized for modeling and generating the synthetic samples of the solar irradiance at the 11$^{\text{th}}$ hour of the day, and another dedicated MC is developed and optimized for modeling and generating the synthetic samples of the solar irradiance at the 18$^{\text{th}}$ hour of the day. This diversification is applied to prices and wind speed, too, even if the wind speed is typically not expected to benefit from this diversification so much, as it does not show a strong daily seasonality.

The grid search for the optimization of the MC architecture is implemented with $o^* = 24$ (to capture information brought by the data observed the previous day at the homologous hour) and $N^* = 200$. A five-state discretization of the grid search is applied to the optimal number of states.

The performance of the proposed methodology is tested against a benchmark (BenchMC) that is a traditional single-order MC with transition probabilities estimated by FFA and with UB. The BenchMC is useful to understand the impact of the alternative SFA for the estimation of the transition probabilities and of the NUB.

A qualitative evaluation of the performance is provided below with Quantile-Quantile (Q-Q) plots [19] of the generated synthetic samples against the observed samples, whereas a quantitative evaluation of the performance is provided in terms of Normalized Mean Absolute Deviations (NMADs) of the mean (μ), standard deviation (σ), and exemplarily the 0.2-quantile ($q^{\langle 0.2\rangle}$) and 0.8-quantile ($q^{\langle 0.8\rangle}$), between the generated synthetic samples and the observed samples. For instance, the NMAD-$q^{\langle 0.2\rangle}$ is calculated as:

$$\text{NMAD-}q^{\langle 0.2\rangle} = \frac{\frac{1}{24}\sum_{h=1}^{24}\left|q_{synth,h}^{\langle 0.2\rangle} - q_{obser,h}^{\langle 0.2\rangle}\right|}{x_{max} - x_{min}}, \tag{15}$$

where $q_{synth,h}^{\langle 0.2\rangle}$ is the 0.2-quantile of the generated synthetic samples related to the h^{th} hour of the day through the entire test window, and $q_{obser,h}^{\langle 0.2\rangle}$ is the 0.2-quantile of the observed samples related to the h^{th} hour of the day.

A. *Case Study*

In this case study, the NUB-L is applied to solar irradiance, whereas the UB is applied to wind speed and electricity prices. MC architectures are developed with either the FFA or the SFA, and with $k = 1$ or $k = 24$. We report the results of the MC architecture optimization in Table I (excluding nighttime hours for solar irradiance, as it is deterministically zero), and three examples of synthetic scenarios (among the 2000 scenarios generated with the optimal structures indicated in Table I) for each of the three variables in Fig. 3.

TABLE I. RESULTS OF THE MC ARCHITECTURE OPTIMIZATION

Hour of the day	**Solar irradiance**		**Wind speed**		**Electricity price**	
	Optimal structure	***Optimal number of states $\bar{N}$***	***Optimal structure***	***Optimal number of states $\bar{N}$***	***Optimal structure***	***Optimal number of states $\bar{N}$***
1	-	-	FFA UB, $\bar{o}$=2, k=1	140	FFA UB, $\bar{o}$=1, k=1	30
2	-	-	FFA UB, $\bar{o}$=2, k=1	40	FFA UB, $\bar{o}$=1, k=1	90
3	-	-	FFA UB, $\bar{o}$=2, k=1	40	FFA UB, $\bar{o}$=1, k=1	90
4	-	-	SFA UB, $\bar{o}$=24, k=1	150	SFA UB, $\bar{o}$=24, k=1	150
5	-	-	FFA UB, $\bar{o}$=2, k=1	30	FFA UB, $\bar{o}$=2, k=1	70
6	SFA NUB-L, $\bar{o}$=24, k=1	15	FFA UB, $\bar{o}$=2, k=1	40	FFA UB, $\bar{o}$=1, k=1	80
7	SFA NUB-L, $\bar{o}$=24, k=1	25	SFA UB, $\bar{o}$=24, k=1	150	FFA UB, $\bar{o}$=1, k=1	20
8	SFA NUB-L, $\bar{o}$=24, k=1	15	SFA UB, $\bar{o}$=24, k=1	130	FFA UB, $\bar{o}$=1, k=1	30
9	FFA NUB-L, $\bar{o}$=2, k=1	15	SFA UB, $\bar{o}$=24, k=1	150	FFA UB, $\bar{o}$=1, k=1	50
10	FFA NUB-L, $\bar{o}$=1, k=1	35	FFA UB, $\bar{o}$=2, k=1	50	FFA UB, $\bar{o}$=1, k=1	40
11	FFA NUB-L, $\bar{o}$=1, k=1	30	SFA UB, $\bar{o}$=24, k=1	150	FFA UB, $\bar{o}$=1, k=1	90
12	FFA NUB-L, $\bar{o}$=1, k=1	40	FFA UB, $\bar{o}$=1, k=1	30	FFA UB, $\bar{o}$=1, k=1	50
13	FFA NUB-L, $\bar{o}$=1, k=1	35	FFA UB, $\bar{o}$=2, k=1	40	FFA UB, $\bar{o}$=1, k=1	80
14	FFA NUB-L, $\bar{o}$=2, k=1	20	FFA UB, $\bar{o}$=1, k=1	110	FFA UB, $\bar{o}$=1, k=1	100
15	FFA NUB-L, $\bar{o}$=1, k=1	35	FFA UB, $\bar{o}$=2, k=1	70	FFA UB, $\bar{o}$=1, k=1	50
16	FFA NUB-L, $\bar{o}$=2, k=1	15	FFA UB, $\bar{o}$=2, k=1	30	FFA UB, $\bar{o}$=1, k=1	40
17	FFA NUB-L, $\bar{o}$=1, k=1	25	FFA UB, $\bar{o}$=2, k=1	30	FFA UB, $\bar{o}$=1, k=1	40
18	SFA NUB-L, $\bar{o}$=24, k=1	15	FFA UB, $\bar{o}$=2, k=1	50	FFA UB, $\bar{o}$=1, k=1	120
19	SFA NUB-L, $\bar{o}$=24, k=1	15	FFA UB, $\bar{o}$=2, k=1	40	FFA UB, $\bar{o}$=1, k=1	130
20	SFA NUB-L, $\bar{o}$=24, k=1	10	FFA UB, $\bar{o}$=2, k=1	40	FFA UB, $\bar{o}$=1, k=1	60
21	-	-	FFA UB, $\bar{o}$=2, k=1	50	FFA UB, $\bar{o}$=1, k=1	90
22	-	-	FFA UB, $\bar{o}$=2, k=1	40	FFA UB, $\bar{o}$=1, k=1	140
23	-	-	FFA UB, $\bar{o}$=1, k=1	30	FFA UB, $\bar{o}$=1, k=1	60
24	-	-	FFA UB, $\bar{o}$=1, k=1	30	SFA UB, $\bar{o}$=24, k=1	10

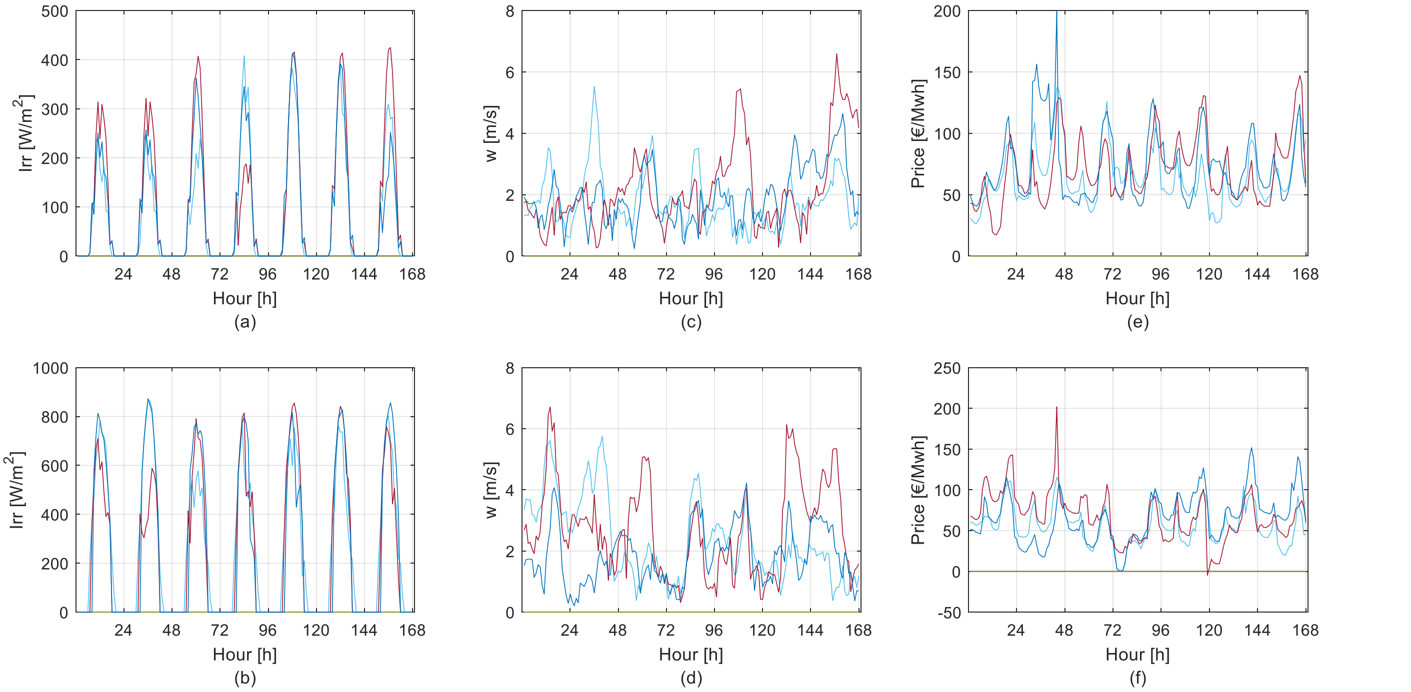

Fig. 3. Three generated synthetic winter-week scenarios of solar irradiance (a), wind speed (c), and electricity price (d). Three generated synthetic summer-week scenarios of solar irradiance (b), wind speed (d), and prices (f).

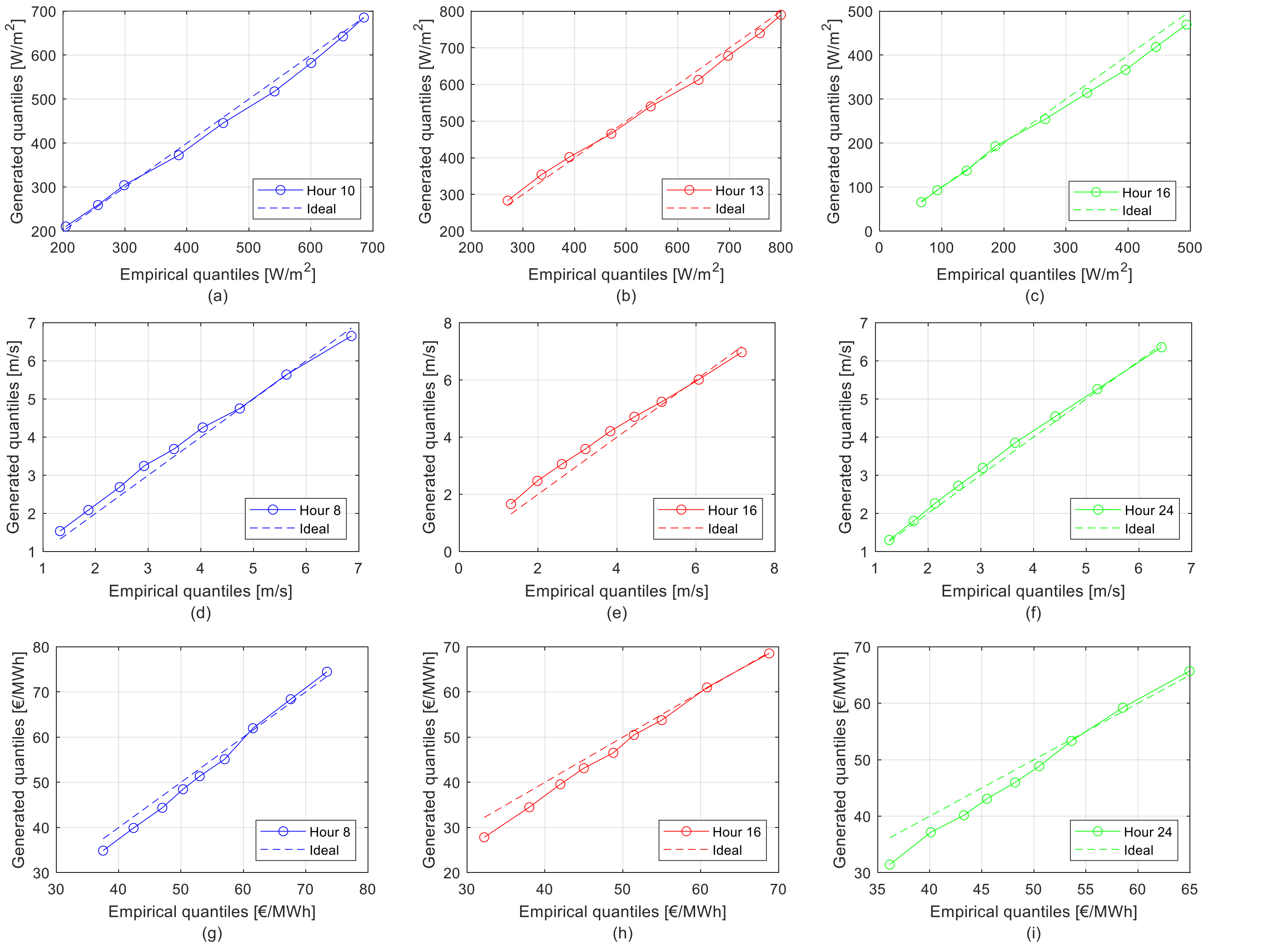

Fig. 4. Q-Q plots of the generated synthetic samples of solar irradiance at 10[th] hour (a), 13[th] hour (b), and 16[th] hour (c) of the day; wind speed at 8[th] hour (d), 16[th] hour (e), and 24[th] hour (f) of the day; electricity prices at 8[th] hour (g), 16[th] hour (h), and 24[th] hour (i) of the day.

From Table I it is possible to note that the SFA is selected six times for solar irradiance (always in correspondence to dawn or dusk hours), and seven times overall for wind speed and price. The FFA is selected in the majority of cases, but only for small MC optimal order $\bar{o}$ (either 1 or 2).

The optimal number of states significantly varies as the hour of the day varies. When the optimal order is $\bar{o} = 24$, the selected approach is always the SFA, denoting a prevalence with respect to the FFA for larger MC orders. The performance of the proposed methodology is qualitatively assessed through the Q-Q plots shown in Fig. 4, and it is quantitatively assessed through the absolute deviation values reported in Table II. From Fig. 4 it appears that the empirical scenarios closely follow the ideal straight line, for all variables and for all the hours of the day. NMAD values in Table II suggest that the proposal improves the representation of the empirical data by up to 71% (for solar irradiance), 67% (for wind speed), and 24% (for prices). In only one case (NMAD-$q^{(0.2)}$, wind speed) the NMAD of the proposal is slightly larger than the NMAD of the BenchMC.

TABLE II. QUANTITATIVE ASSESSMENT OF THE PERFORMANCE THROUGH NORMALIZED MEAN ABSOLUTE DEVIATION INDICES

Variable	Method	Metric			
		NMAD-μ	***NMAD-σ***	***NMAD-$q^{(0.2)}$***	***NMAD-$q^{(0.8)}$***
Solar irradiance	Proposal	0.0064	0.0045	0.0029	0.0145
	BenchMC	0.0204	0.0098	0.0100	0.0304
Wind speed	Proposal	0.0069	0.0117	0.0158	0.0036
	BenchMC	0.0135	0.0183	0.0153	0.0110
Electricity price	Proposal	0.0136	0.0122	0.0203	0.0046
	BenchMC	0.0168	0.0131	0.0208	0.0060

V. CONCLUSION

This paper presents a generalized methodology for the generation of synthetic scenarios of heterogeneous random variables involved in road tunnel planning problems. The proposed methodology is unified and applied for the generation of synthetic scenarios of the different and heterogeneous random variables that are involved in the decision-making process for APROT. The methodology is based on MCs that are developed under different approaches (FFA and SFA) for the estimation of the transition probabilities, and with different procedures (UB and NUB) for the identification of bins. This allows for dealing with variables of different natures and sources, as it occurs in APROT planning.

Numerical applications based on actual data in an Italian case study confirm the suitability of the proposal in generating reliable and representative synthetic scenarios for long-term horizons, outperforming a MC-based benchmark by up to 71%.

Future works might involve taking into account the possible variable interdependence in the generation of synthetic scenarios (e.g., through multivariate or correlation-based approaches).

REFERENCES

[1] G. Uddin, et al., "Microgrids: A review, outstanding issues and future trends," Energy Strat. Rev., vol. 49, 101127, 2023.

[2] A. Bracale, et al., "Models and Methods for Sustainable Road Tunnels Operating as Microgrid and RES Energy Community", in proc. 2024 IEEE Int. Conf. Environm. Electr. Eng. / 2024 IEEE Ind. Comm. Power Syst. Europe (EEEIC/I&CPS Europe), 17-20 June 2024, Rome (Italy).

[3] F. Agha Kassab, R. Rodriguez, B. Celik, F. Locment, and M. Sechilariu, "Comprehensive Review of Sizing and Energy Management Strategies for Optimal Planning of Microgrids with PV and Other Renewable Integration," Appl. Sci., vol. 14, 10479, 2024.

[4] J. Li, J. Zhou, and B. Chen, "Review of wind power scenario generation methods for optimal operation of renewable energy systems," Appl. Energy, vol. 280, 115992, 2020.

[5] A. Shamshad, M.A. Bawadi, W.M.A. Wan Hussin, T.A. Majid, and S.A.M. Sanusi, "First and second order Markov chain models for synthetic generation of wind speed time series", Energy, vol. 30, pp. 693-708, 2005.

[6] G. D'Amico, F. Petroni, and F. Prattico, "First and second order semi-Markov chains for wind speed modeling," Physica A: Stat. Mech. Appl., vol. 392, no. 5, pp. 1194-1201, 2013.

[7] V. Di Giorgio, R. Langella, A. Testa, S.Z. Djokic, and M. Zou, "First order non-homogeneous Markov chain model for generation of wind speed and direction synthetic time series," in proc. 2020 IEEE Int. Conf. Probab. Methods Appl. Power Syst. (PMAPS), 2020.

[8] V. Di Giorgio, A. Testa, M. Zou, R. Langella, and S.Z. Djokic, "Simplified Markov Chain Models for Generation of Synthetic Time Series of Wind Speed and Direction," in proc. 2022 IEEE Int. Conf. Probab. Methods Appl. Power Syst. (PMAPS), 2022.

[9] S. Eberle, D. Cevasco, M.-A. Schwarzkopf, M. Hollm, and R. Seifried, "Multivariate simulation of offshore weather time series: a comparison between Markov chain, autoregressive, and long short-term memory models," Wind, vol. 2, no. 2, pp. 394-414, 2022.

[10] S. Miao, G. Ning, Y. Gu, J. Yan, and B. Ma, "Markov Chain model for solar farm generation and its application to generation performance evaluation," J. Clean. Prod., vol. 186, pp. 905-917, 2018.

[11] Â. Frimane, T. Soubdhan, J.M. Bright, M. Aggour, "Nonparametric Bayesian-based recognition of solar irradiance conditions: Application to the generation of high temporal resolution synthetic solar irradiance data," Solar Energy, vol. 182, pp. 462-479, 2019.

[12] M. Rayati, P. De Falco, D. Proto, M. Bozorg, M. Carpita, "Generation Data of Synthetic High Frequency Solar Irradiance for Data-Driven Decision-Making in Electrical Distribution Grids," Energies, vol. 14, 2021.

[13] J. M. Bright (ed), "Synthetic solar irradiance: Modeling solar data," AIP Publishing, 2021.

[14] A. Bracale, et al., "Optimal planning of distributed energy resources for sustainable road tunnels under uncertainty," under review at Sustain. Energy Grids Netw., 2025.

[15] A. Bracale, G. Carpinelli, P. De Falco, "A Probabilistic Competitive Ensemble Method for Short-Term Photovoltaic Power Forecasting," IEEE Trans. Sustain. Energy, vol. 8, no. 2, pp. 551-560, 2017.

[16] W.-K. Ching, X. Huang, M.K. Ng, T-K. Siu, "Markov Chains: Models, algorithms and applications," (2nd ed.), Springer Nature, 2013.

[17] European Centre for Medium-range Weather Forecasts. Available online: https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5

[18] European Network of Transmission System Operators website. Available online: https://www.entsoe.eu

[19] J.I. Marden, "Positions and QQ plots," Statist. Sci., vol. 19, no. 4, pp. 606–614, 2004.