A Bayesian Hierarchical Model to Create Synthetic Power Distribution Systems

Henrique O. Caetano, Luiz Desuó N.,
Matheus de S. S. Fogliatto, Carlos D. Maciel
Department of Electrical and Computing Engineering
University of São Paulo (EESC/USP) - São Carlos, Brazil
{henriquecaetano1}@usp.br

Vitor P. Ribeiro, José A. P. Balestieri,
Faculty of Engineering and Science
São Paulo State University (UNESP) - Guaratinguetá, Brazil
{vitor.p.ribeiro}@unesp.br

***Abstract*—The growing complexity of Power Distribution Systems, driven by distributed generation, renewable energy integration, and increasing demand, has led to restricted access to DS data due to security and privacy concerns. This study addresses limited data accessibility by proposing a hybrid approach for crafting synthetic power distribution systems tailored for power system analysis and control. Synthetic power distribution systems refer to artificially generated models that faithfully replicate real-world DS features while upholding security and privacy constraints. This innovative methodology merges a Bayesian Hierarchical Model with Markov Chain Monte Carlo techniques, utilizing georeferenced data to capture intricate system dependencies, feeder configurations, switch statuses, and load node distributions. Leveraging OpenStreetMaps for DS topology, the approach incorporates expert knowledge and real-world data. Results highlight the methodology's ability to evaluate credible intervals for parameters, facilitating a probabilistic assessment of uncertainties and enhancing decision support in power system analysis and control. Findings affirm the hybrid approach's efficacy in generating realistic synthetic DSs, bridging the gap between statistical and georeferenced methodologies for advanced power system analysis and control. The capacity to generate synthetic DSs provides valuable insights into power system dynamics, addressing security, privacy, and data accessibility concerns for a more informed decision-making process.**

***Index Terms*—distribution systems, synthetic test cases, bayesian hierarchical model, georeferenced data**

I. Introduction

Distributed generation, renewable energy resources, and the ever-growing demand for electricity have collectively contributed to the increasing complexity of power distribution systems (DS) [1]. However, due to security concerns, data related to DS are often restricted, posing a challenge for researchers and engineers who require access to reliable datasets for training and testing power system analysis and control algorithms [2]. Consequently, there is an urgent need for methodologies that can effectively generate synthetic DS from publicly available data, bridging the gap between data scarcity and the demand for data-driven solutions in the power distribution domain.

Previous research has primarily focused on the development of tools based on open data sources, such as OpenStreetMap (OSM) [3]. These georeferenced approaches utilize street location and distribution information to define the general topology of electrical systems, providing an initial framework for synthetic DS generation [1]. However, these methods still face challenges in specifying electrical parameters and network demands accurately, often resorting to simplified assumptions, such as the proportionality of load demand to building size, which may not fully capture the intricacies of real-world distribution systems [1]. On the other hand, statistical tools have been utilized to define probability distribution functions for electrical properties of DS, allowing for the generation of synthetic systems that reflect statistical characteristics of real data [2]. However, these purely statistical methods may lack georeferenced information and fail to produce realistic representations of urban DS with accurate spatial and topological features [2].

To address these limitations and create more accurate synthetic DS, we propose a novel methodology that integrates statistical and georeferenced approaches, presenting a hybrid model for synthetic power distribution systems. The Bayesian Hierarchical Model (BHM) is central to our approach, a powerful statistical method known for its ability to manage intricate dependencies and variable interactions [4]. BHMs offer the advantage of capturing the correlations and interactions between various power distribution system components, including transformers, feeders, and substations, leading to more comprehensive and realistic representations of the DS.

Our proposed BHM leverages the Markov Chain Monte Carlo (MCMC) method to sample elements of the synthetic DS from the posterior distribution. By combining expert knowledge and real-world data, we construct credible intervals (CI) to derive posterior probabilities of parameters, allowing for a robust and data-driven learning process. One significant contribution of our approach is the evaluation of several parameters as random variables, each with its own highest density interval (HDI), instead of single point estimates. This unique characteristic enables our model to effectively assess the impact of uncertainties and quantify the importance of each component in the synthetic system in a probabilistic framework, enhancing decision support capabilities. This contrasts

Submitted to the 22nd Power Systems Computation Conference (PSCC 2024).

with conventional single point estimates found in the literature, which may not fully account for the inherent uncertainties in power distribution systems. Our work not only provides a more accurate and versatile representation of synthetic DS but also offers insights into the integration of statistical and georeferenced data, bridging the gap between two distinct methodologies for synthetic DS generation.

The rest of this article is structured as follows: Section II provides a concise literature review of recent works related to the creation of synthetic power distribution systems. In Section III, we define the methodology of our proposed work, which incorporates a hybrid model integrating georeferenced data and statistical methods. Section IV presents the key findings of our study, and finally, Section V offers concluding remarks.

II. Literature Review

The process of creating synthetic distribution systems has been in the literature in various forms. The interest in such approaches comes down to two main factors: Georeferenced data is usually open available from platforms such as the OpenStreetMaps [3]; Electrical data from power distribution systems are usually not publicy available due to security reasons [5]. In this scenario, two main approaches are considered: either using georeferenced data directly to create the full synthetic distribution system; or using statistical tools to fit real data from real distribution systems to create fully synthetic distribution systems.

In the exploration of creating synthetic distribution systems, the literature reveals two predominant strategies influenced by the dichotomy between the open availability of georeferenced data and the restricted access to electrical data due to security concerns [5]. The first strategy involves the direct utilization of georeferenced data to shape the synthetic network, exemplified by works such as [1], [6], [7], [8], [9]. However, a common drawback in these approaches lies in the oversimplification of the load estimation process, assuming a proportional relationship between the active power of each node and the size of nearby residences or buildings. This simplification, identified as a potential limitation, raises accuracy concerns, prompting the necessity for additional data sources, including statistical information, to refine load estimation precision. Furthermore, recognized gaps in the allocation of switches and consideration of electrical characteristics within current georeferenced approaches underscore the need for future research to address these critical aspects.

Conversely, a distinct set of works opts for statistical approaches, fitting real distribution system data and generating synthetic systems through sampling from statistical models, as evidenced by studies like [10] and [2]. While these statistical methodologies demonstrate efficacy in modeling random variables such as active power and feeder distances, a notable limitation surfaces—they do not incorporate georeferenced data as a foundational element for synthetic network generation. Consequently, the resultant systems lack geospatial representation, limiting their suitability, particularly in modeling distribution systems within urban areas.

It is clear that the existing literature treats statistical and georeferenced approaches to synthetic distribution system creation as distinct methodologies, each with its limitations. The identification of these limitations becomes more pronounced when considering the methods separately. However, it is evident that a synthesis of these approaches holds significant potential for overcoming individual drawbacks. Integrating statistical models with georeferenced data not only enables the creation of more accurate and geospatially representative synthetic power distribution systems but also paves the way for advancing the field in addressing urban-specific challenges.

III. Materials and Methods

The primary objective of this study is to employ a BHM or the allocation and creation of synthetic distribution systems, effectively bridging the gap between georeferenced data and statistical knowledge, including the mean, variance, and credible interval of key variables, such as power demand and load deviation, which can be extracted from real-world data.. This approach aims to address the need for a methodology that takes into account both spatially referenced data and statistical insights. Therefore, this section will be divided into two parts: Section III-A demonstrates the utilization of georeferenced data from a Brazilian city to establish essential topological parameters relevant to the distribution system. Conversely, Section III-B showcases how real-world data from power distribution systems is integrated with a BHM to generate the final parameters for the power distribution system. Figure 1 shows a complete flowchart of the proposed methodology, where both the aforementioned steps are detailed, making reference of each procedure that are detailed hereinafter.

The python Pymc4 library [11] was used to model the BHM. The PandaPower package [12], an open source tool for power system modelling, was used to model the synthetic power distribution system, as well as its individual parameters. The Networkx library [13] was used to calculate graph properties, such as distance from elements in the distribution system.

A. Using georeferenced data

This study involves the development of a synthetic power distribution system originating from a real urban area. The chosen urban area for this investigation is São Paulo, Brazil. Utilizing georeferenced data accessible on OpenStreetMaps [3], Figure 2 illustrates the geographical context of the city. This public available data gives information about the road network in the form of a Directed Graph, where edges redefine the streets, and the nodes constitute the intersection between each lane.

Georeferenced data for substation locations can be accessed through OpenStreetMaps, resulting in a graph representing the city with nodes and edges. Following this, establishing the positions of feeders that depart from these substations involves several sequential steps:

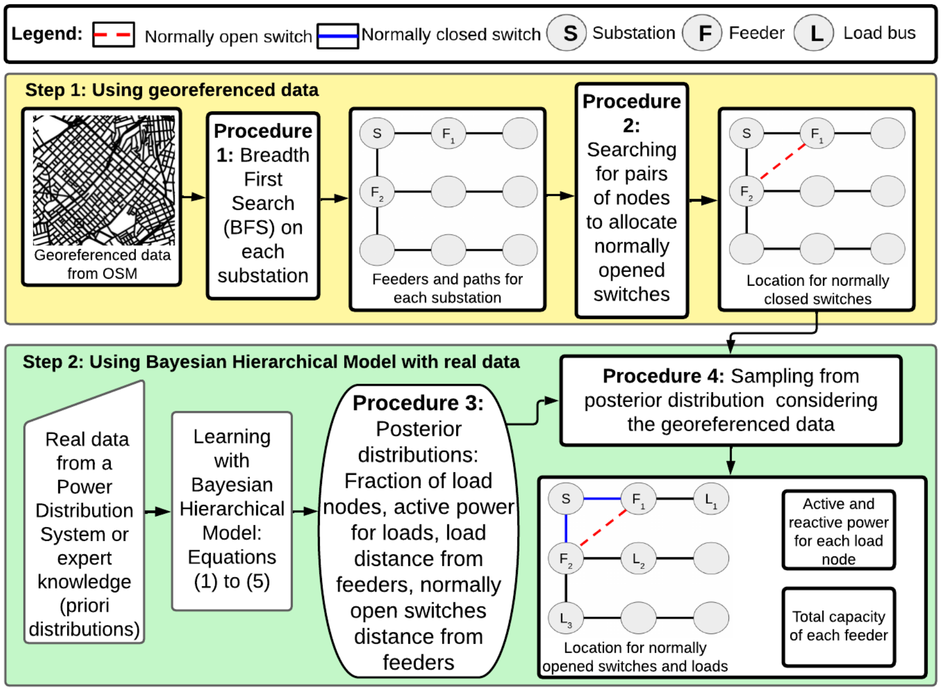

Legend:
Normally open switch
Normally closed switch
S Substation
F Feeder
L Load bus
Step 1: Using georeferenced data
Georeferenced data from OSM
Procedure 1: Breadth First Search (BFS) on each substation
S
F1
F2
Feeders and paths for each substation
Procedure 2: Searching for pairs of nodes to allocate normally opened switches
Location for normally closed switches
Step 2: Using Bayesian Hierarchical Model with real data
Real data from a Power Distribution System or expert knowledge (priori distributions)
Learning with Bayesian Hierarchical Model: Equations (1) to (5)
Procedure 3: Posterior distributions: Fraction of load nodes, active power for loads, load distance from feeders, normally open switches distance from feeders
Procedure 4: Sampling from posterior distribution considering the georeferenced data
L1
L2
L3
Location for normally opened switches and loads
Active and reactive power for each load node
Total capacity of each feeder

Fig. 1. Flowchart of the proposed methodology. In Step 1, the georeferenced data is used to allocate the feeders and electrical paths of each substation (Procedure 1) and normally closed switches (Procedure 2). In Step 2, real data or expert knowledge can be used to fit the BHM, calcualte the posterior distribution (Procedure 3) and sample from posterior distribution to allocate the reamining elements of the synthetic distribution system (Procedure 4).

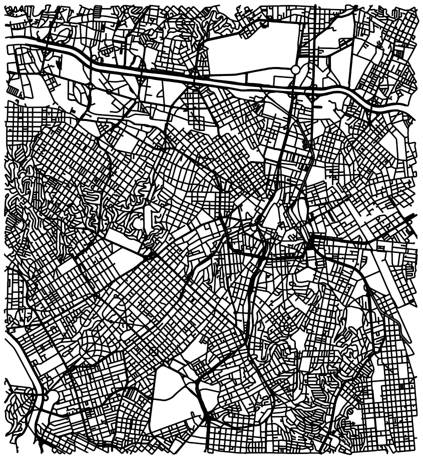


Fig. 2. Road Network from the city of São Paulo, Brazil. The data is taken from OpenStreetMaps [3], and gives the road network in the form of a Directed Graph. This georeferenced data is used as a starting point to create the final synthetic distribution system.

Procedure 1

1) For each substation ($s \in S$) in the georeferenced data, the road network serves as a starting point for conducting a Breadth First Search, originating from the substation's location. The result of this process is a set of paths (P) leaving from the original substation;
2) Considering each path in the path set ($p \in P$), and the graph created to represent the city, the closest node to the original substation is determined, and a feeder is allocated at that node.
3) The active power of each allocated feeder is determined by summing the active power of all load buses connected to it. This ensures that the allocated feeder can meet the demand under normal operational conditions. The specific active power of each load bus will be further defined in Section III-B.

The aforementioned procedure is suitable for defining the proper location of the feeders in the original georeferenced data. The next step involves the allocation of normally opened switches in the network. These switches facilitate the transfer of active power between nodes. The switch allocation procedure is as follows (for each pair of paths):

Procedure 2

1) For each pair of paths leaving the original feeder ($p_1, p_2 \in P$), the closest pair of nodes (one from each path) is identified based on their distance.
2) It is then verified whether each feeder can accommodate the combined demand of both downstream sections in case a power transfer is required. If so, the switch is allocated at that location.
3) The procedure is repeated for the next closest pair of nodes (again, one from each path) until all suitable switch locations are determined.

By adhering to these steps, the optimization of feeder and switch allocation (meaning their exact position in the final synthetic network) is achieved by ensuring that the sectors formed by the placement of normally opened switches can adequately support the demand, a crucial aspect during system failures. This process significantly enhances the efficiency and reliability of the distribution system.

B. Using Bayesian Hierarchical Model with real data

After the procedures described in Section III-A, the locations of substations and feeders are generated. In this Section, the remaining parameters for the power distribution system will be modelled by using a BHM. Particularly, the two main remaining components are as follows:

- Location of each load bus (as well as its active and reactive power);
- Location of normally closed switches (circuit breakers) which are used to disconnect sectors in case of failure)

The primary advantage of the BHM, in this context, lies in its ability to incorporate multiple layers of random variables. In this modeling approach, variables at the highest level are considered first, and variables at lower levels take the distribution of the higher-level variables as a reference. This hierarchical structure allows for a more comprehensive and flexible representation of the data, capturing dependencies and relationships between different variables effectively.

Equations 1 to 7 define the complete BHM proposed in this study. In Equation (1), a mixture model is introduced, where the probability distribution of the random variable x is determined by the same likelihood distribution - $f(x)$ - but each with its individual weight contribution (w_i) and hyperparameters θ_i. This mixture model is applied to two variables of interest: the distance from load nodes to feeder (d_{LN}) and the distance from normally closed switches from feeder d_{NC}. In both cases, we use the Poisson distribution as the preferred likelihood function, as shown in Equation (2), where λ_i is the expected rate of occurrences. Mixture Poisson models have been successfully employed to redefine heterogeneous characteristics of power systems [14], and as demonstrated in Section IV, the real-world data used to fit our model aligns well with this choice.

$$f(x|w,\theta) = \sum_{i=1}^{N} w_i f(x|\theta_i), \forall i \in \{1,2\}, \forall x \in \{d_{LN}, d_{NC}\} \tag{1}$$

$$f(x|\lambda_i) = \frac{e^{-\lambda_i}\lambda_i^x}{x!} \forall i \in \{1,2\}, \forall x \in \{d_{LN}, d_{NC}\} \tag{2}$$

After defining the location of each load bus in the distribution system, the next step is to define its active and reactive power. In this sense, similar to the work proposed in [15], the process is divided in three main random variables: fraction of load nodes in the power distribution system as a whole (F_L) (where non-load nodes are considered as connection buses), modelled as a beta distribution function as shown in Equation (3), where a (or α) and b (or β) are the parameters of the distribution; load deviation (D_l), modelled as a t-location scale distribution as shown in Equation (4), where ν,μ and σ are the shape, location and scale parameters respectively; power factor for the whole network (PF), modelled as a simple three possibilities as shown in Equation (5), where u is a random uniform distribution in the closed interval $[0, 1]$. This parameterization, as previously employed in the literature [1], can essentially be interpreted as follows: The protective feature (PF) has a 16.49% probability of being equal to 0.85, followed by a $(0.27 - 0.1649) \cdot 100 = 10.51\%$ probability of being equal to 0.90, and finally, a $100 - 16.49 - 10.51 = 73\%$ probability of being equal to 0.95. Depending on the specific distribution system (DS) being modeled and relying on either expert knowledge or real data, this parameterization can be extended to accommodate additional PF possibilities, each with its respective set of probabilities.

$$f(F_L|a,b) = \frac{\Gamma(a+b)x^{a-1}(1-x)^{b-1}}{\Gamma(a)\Gamma(b)} \tag{3}$$

$$f(D_l|\nu,\mu,\sigma) = \frac{\Gamma(\frac{\nu+1}{2})}{\sigma\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}[\frac{v+(\frac{x-\mu}{\sigma})^2}{\nu}]^{-\frac{\nu+1}{2}}, \forall l \in F_L \tag{4}$$

$$PF = \begin{cases} 0.85, & \text{if } 0 < u \leq 0.1649 \\ 0.90, & \text{if } 0.1649 < u \leq 0.27 \\ 0.95, & \text{otherwise} \end{cases} \tag{5}$$

Finally, additional parameters can be calculated based on the aforementioned random variables. First, the active power for each load node ($P_l, \forall l \in F_L$) can be calculated based on Equation (6), where μ_P is the mean active power (which can be defined based in either previous data or expert knowledge). Additionally, the reactive power can be calculated from Equation (7), considering the power factor for the whole network and the active power for each individual load node.

$$P_l = \mu_P + D_l \cdot \mu_P, \forall l \in F_L \tag{6}$$

$$Q_l = P_l \cdot tan(arccos(PF)), \forall l \in F_L \tag{7}$$

After properly modelling the variables of interest as random variables, with their respective distribution functions, the next step is to define the value of each hyperparameter of the chosen distributions. In order to do so, either expert knowledge or real data can be used. Given such data, the following procedure is used to fit the data into the BHM:

Procedure 3

1) For each feeder in the network, calculate the fraction of load nodes (given all nodes). This will be used into Equation (3);
2) For each load node in the network, calculate the distance from it to its respective feeder. Use this data to fit into Equation (1) with $x = d_{LN}$;

3) For each normally closed switch in the network, calculate the distance from it to its respective feeder. Use this data to fit into Equation (1) with $x = d_{NC}$;
4) Calculate the mean active power for the system as a whole. Store it in μ_P. For each load node in the network, calculate the difference between its actual power and the mean ($D_l = P_l - \mu_P$). Use it to fit into equation (4)

Procedure 3 enables the definition of the posterior distribution for each parameter of interest. This essentially involves fitting the statistical model to real data, allowing these distributions to serve as a basis for sampling data for new systems. Although synthetic in nature, these new systems retain characteristics derived from the data of a real system.

The subsequent and crucial step involves utilizing the posterior distribution of the real-world data to generate the synthetic distribution system. However, due to potential significant variations between the distribution system where the data was collected and the georeferenced location where the synthetic distribution system will be created (including differences in the number of loads, substations, and the active power demand of each load node), a normalization procedure is necessary. This normalization procedure will be described as Procedure 4.

Procedure 4

1) Initially, define as a parameter: the number of normally closed switches to be allocated: N_{NC};
2) Calculate the mean path length distance for all feeders in the georeferenced data: d_{geo}; calculate the mean path length distance for all feeders in the original real data (or define it through expert knowledge): d_{real}
3) Sample d_{NC} from equation (1). Try to find a edge from the georeferenced data whose distance from feeder is equal to $d_{NC} \cdot \frac{d_{geo}}{d_{real}}$ (normalization procedure that takes into account possible distinct scales between real data and the synthetic system). Allocate a normally closed switch in such a node. Repeat this procedure for N_{NC} times;
4) Given the partial distribution system with feeder locations generated in Section III-A: calculate the total number of nodes (N_t). Sample F_L from Equation (3). Consider $N_L = F_L \cdot N_t$ as the normalized number of load noads;
5) Sample d_{LN} from equation (1). Try to find a node from the georeferenced data whose distance from feeder is equal to d_{LN}. Allocate a load in such a node. Repeat this procedure for N_t times;
6) Define the PF of the noad using Equation (5);
7) Define μ_P that is desired for the power distribution system. For each allocated load node, sample its deviation from Equation (4), and use Equations (6) and (7) to define its active and reactive power
8) Finally, define the demand of each feeder of the network as the total sum of active power from all load nodes connected to it.

After this final procedure, the output of the model is a synthetic distribution system with normally opened and closed switches; load nodes with active and reactive demand and substations with connected feeders.

IV. Results and Discussions

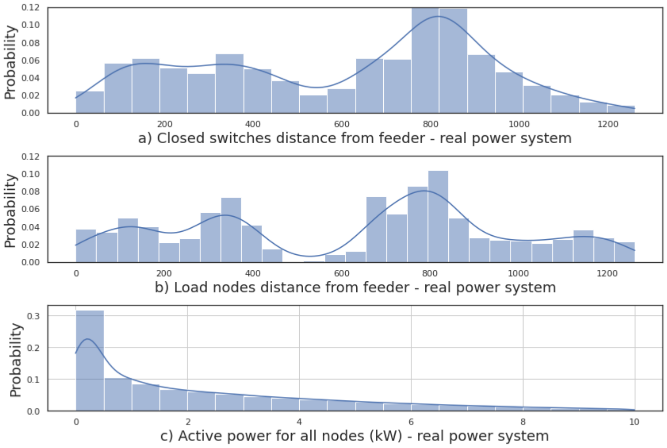
The results are organized into sections to enhance reader comprehension. In Section IV-A, we present data from a real Distribution System (DS) and justify its use. Section IV-B showcases the posterior distribution of the Bayesian Hierarchical Model (BHM) fitted into real data, including its parameters and visual representation. Moving on to Section IV-C, we provide an example of a synthetic distribution system created using the proposed methodology, along with a validation of the main results. Finally, in Section IV-D, we present the computational time associated with the proposed methodology.

A. *Real Data*

In this work, in order to demonstrate the effectiveness of the BHM to deal with real world data, a Brazilian power distribution system is used. For more details regarding the real data applied, please refer to the work done in [16]. The system has over 40000 buses, 3800 switches and 36000 loads. In this power system, the connections between buses, feeders and substations are well-known, as well as the active power of each load node.

To validate our hypothesis of applying a mixture model for the hop distance from switches to feeders, the hop distance was calculated for each bus in the Brazilian DS. A histogram was created, as shown in Figure 3 a). While there is a peak frequency at around 800, frequencies around 200 and 400 also appear at a considerable ratio, which seems to indicate that a single distribution is not enough to model the variable's behaviour, thus serving as a motivation to model the maximum hop distance from feeders as a Mixture Poisson Model. In Figure 3 b), the same procedure is done with load nodes distance. It is clear that a mixture behavior is also presented in such variable, which was also modelled as a mixture poisson. Both closed switches distance and load nodes distance are represented in Equation 1. Moreover, the lower values observed in the central part of Figure 3 a) and b) can be attributed to the dual purposes of closed switches. They are strategically positioned: either in close proximity to the feeder, facilitating disconnection from the rest of the network, or at a relatively greater distance to isolate loads in the event of disruptions. As a result, middle distances are less common, with a higher concentration towards the tails of the distribution.

Finally, in Figure 3 c), we define the distribution of active power for each load node in the real data. The plot reveals that this variable predominantly consists of values concentrated at low active power levels (close to 0), with the probability decreasing as active power increases. Due to the nature of this variable, which can only have positive values, and the heterogeneous distribution observed in the real data, the procedure outlined in Section III-B and adapted from [15] is well-suited. This adaptation is essential because the t-location scale distribution is useful for modeling data distributions with heavier tails (more prone to outliers) than the normal distribution, and can become a normal distribution depending on the values of its parameters. Such adaptability to real-world data is crucial for the BHM to effectively fit the data.


a) Closed switches distance from feeder - real power system

b) Load nodes distance from feeder - real power system

c) Active power for all nodes (kW) - real power system

Fig. 3. Histogram for the real dataset obtained from a Brazilian Power Distribution System, serving as the basis for fitting the Bayesian Hierarchical Model (BHM) proposed in this work. The data from (a) and (b) were fitted to a mixture Poisson model, while the data in (c) was fitted to a t location-scale distribution. In this representation, distance refers to the number of edges in the shortest path connecting each element from the feeder. The blue curve overlaid on the histogram represents the kernel density estimate (KDE) plot, enhancing the visual understanding of the distribution of observations within the discrete histogram.

B. *Fitting real data into BHM*

After exploring the real data from a Brazilian power distribution system, procedure 3 detailed in Section III is used to properly fit the data. Table I shows the final values of hyperparameters for the Fraction of Load Nodes and Load Deviation. Table II shows the hyperparameters for the mixture model used to fit the distance from normally closed switches to feeder (d_{LN}). Table III shows the mixture model for nodes with power (d_{LN}). For the mixture Poisson models, the Highest Density Interval (HDI) is also shown, since this is the output of a BHM, and as such, each parameter is considered as a random variable.

By comparing the mean values presented in Table II with the histogram from real data in Figure 3 a), it is possible to observe that the mixture Poisson properly fitted the real data into the mixture poisson model. The first Poisson distribution has a much higher mean value (685.60 for λ_0), with a higher weight mean (0.78 for w_0). The second Poisson distribution has a mean value of 56.05 (λ_1), but with a much higher standard deviation of 99.78. This type of result is coherent with the real data, where a majority of switches have a large distance from feeder (between 600 and 1000), while a smaller (but significant) number of switches have its distance between 0 and 500.

A similar interpretation can be done regarding the results presented in Table III. In this case, once again, the first poisson curve has a higher value for the mean of the poisson hyperparameter (λ_0) as well as its weight (w_0). This is coherent with real data presented in Figure 3 b), where the distance from load nodes to feeder can be splitted into two main groups: between 0 and 500 (which was fitted into λ_1 and from 500 onwards (which was fitted into λ_0).

The significance of employing a BHM becomes evident in both the presented mixture Poisson models, as detailed in Tables II and III. The inclusion of Highest Density Interval (HDI) calculations for each parameter, including weights, proves crucial. This not only enhances our understanding of the fitted data but also provides a range of intervals for utilization in constructing the synthetic distribution system. In contrast to previous approaches that utilized a single static value for each parameter, our method offers a more nuanced and adaptable approach. By incorporating the BHM, we gain the flexibility to explore diverse intervals, such as the mixture poisson to model distributions with heterogeneous characteristics, thereby refining the precision of the synthetic distribution system and offering a more comprehensive representation of the underlying data characteristics.

TABLE I
Distribution curves (and their respective parameters) fitted to historical data from a Brazilian power distribution system, regarding the active power of each load node.

Property	Distribution	Parameters
Fraction of Load Nodes (F_L)	Beta	$\alpha = 3.03, \beta = 49.54$
Load Deviation (D_l)	tLocationScale	$\mu = -0.001, \sigma = 0.002, \nu = 1.46$

TABLE II
Means (μ), Standard deviations (σ), and their Highest Density Intervals (HDI) for the Mixture Poisson used to model the distance from normally closed switches to feeders (d_{NC}). The data of a Brazilian power distribution system was used to fit the model.

Parameter	μ	σ	**HDI** (3%)	**HDI**(97%)
λ_0	685.60	72.95	651.15	825.54
λ_1	56.05	99.78	0.73	250.64
w_0	0.78	0.06	0.62	0.83
w_1	0.22	0.06	0.17	0.38

TABLE III
Means (μ), Standard deviations (σ), and their Highest Density Intervals (HDI) for the Mixture Poisson used to model the distance from load nodes to feeder (d_{LN}). The data of a Brazilian power distribution system was used to fit the model.

Parameter	μ	σ	**HDI** (3%)	**HDI**(97%)
λ_0	672.291	113.766	401.471	754.581
λ_1	24.115	28.444	0.552	65.631
w_0	0.485	0.131	0.257	0.627
w_1	0.515	0.131	0.373	0.743

C. *Creating a synthetic system from data*

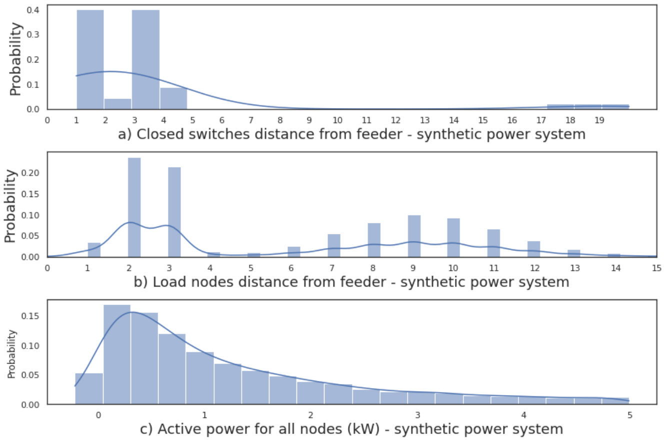
To demonstrate the effectiveness of our proposed approach in creating synthetic power distribution systems based on the posterior distribution of real data, Figure 4 illustrates the histogram for a single sample of the synthetic distribution system constructed using Procedure 4, as explained in Section III-B. For this particular case, we fixed the number of normally closed switches at $N_{NC} = 50$. It is essential to note that the main path length of each feeder in the network is considerably smaller than that of the original real data used to fit the BHM. Consequently, the values of distances for the real data (Figure 3 a) and b)) are significantly smaller compared to the synthetic distribution system (Figure 4). This discrepancy is expected and results from the normalization technique detailed in Procedure 4. The normalization procedure enables the creation of the synthetic distribution system based on georeferenced dgata of varying sizes, even if it differs from the real data used to derive the posterior distributions.

Conversely, Figures 3 c) and 4 c) present comparable shapes and scales. The congruence arises from the absence of normalization in the active power for each node within our proposed approach. This shape similarity underscores the efficacy of our method in drawing samples from the posterior distribution. Yet, nuanced differences emerge, inherent to the Bayesian nature of our model. The sampling process considers the posterior distribution of each parameter rather than a singular, static value. Consequently, even with a substantial number of samples, each synthetic distribution sample exhibits relative distinctions from the actual distribution of real data. Further exploration into quantifying this dissimilarity geometrically or mathematically would provide a more precise measurement.

To address the potential concern of dissimilarity between synthetic and real data histograms, it's crucial to note that a geometrical or mathematical measure may not be ideal in this case. The synthetic data represents a small sample from the posterior distribution learned from real data. Therefore, the final histogram is expected to exhibit dissimilarity due to the inherent variability in the Bayesian framework. With an increased number of samples, the synthetic distribution will undoubtedly converge to the distribution fitted from real data, illustrating the robustness of our approach in capturing the underlying dynamics of the power distribution system.

It is crucial to highlight that the histogram presented in Figure 4 represents only one instance of a synthetic system. As per our proposed methodology, several different synthetic systems can be created based on the same georeferenced data. This approach is distinct from previous methods, where a single synthetic system is generated for each georeferenced data. The ability to create multiple synthetic systems from the same georeferenced data is made possible by our novel hybrid statistical and georeferenced approach for the creation of synthetic power distribution systems, which is the first of its kind in the literature.

Emphasis should be placed on the resultant synthetic grid and its corresponding parameters. The generated system consists of 12.538 buses, 25.436 lines, and 9,813 loads, characterized by a mean active power of 22.06mW and reactive power of 8.37MVar. Additionally, the system features 50 normally closed switches and 183 normally opened switches. All elements within this synthetic grid are georeferenced, possessing a well-defined geographical position, as the original urban area considered is georeferenced. Importantly, the georeferencing aspect equips the system for simulating problems requiring component location, such as Fault Location Isolation and Service Restoration (FLISR). This capability enhances the practical applicability of the synthetic grid in diverse scenarios.



a) Closed switches distance from feeder - synthetic power system

b) Load nodes distance from feeder - synthetic power system

c) Active power for all nodes (kW) - synthetic power system

Fig. 4. Histogram for a single synthetic distribution system created by sampling from the posterior distribution of the final BHM. In a) and b), a mixture Poisson model was used. In c), t location-scale distribution is considered. The procedure for sampling from posterior distribution is presented in Section III-B. The blue curve overlaid on the histogram represents the kernel density estimate (KDE) plot, enhancing the visual understanding of the distribution of observations within the discrete histogram.

Finally, it is essential to clarify that our approach is not primarily designed for the quantitative prediction of loads or load profiles. Quantitative evaluation in this context falls outside the intended scope of our methodology. Instead, the primary objective is to leverage real data to learn from the underlying distribution, subsequently facilitating the generation of synthetic systems sampled from the posterior distribution learned from the data. While a direct quantitative assessment of load predictions may not be applicable, it can be reasonably asserted that the generated load profiles will exhibit similar behavioral characteristics to the real data from which the posterior distribution was derived. This aligns with the fundamental purpose of our approach, which is to faithfully capture and replicate essential features of real-world power distribution systems.

D. *Computational Time*

The computational performance of our proposed method holds significant implications for its practical utility. Executed on a personal computer with 8GB of RAM and an Intel Core i7-8550U CPU @ 1.80GHz, utilizing Python 3.8.10, all simulations' computational times for each step are detailed in Table IV. It is noteworthy that Procedure 4, involving the synthesis of distribution systems by sampling from the posterior distribution, stands out for its exceptional efficiency, requiring only 10.8 seconds. This efficiency underscores the method's practical feasibility for generating multiple synthetic distribution systems. Leveraging Procedure 4 iteratively with distinct inputs enables the creation of diverse synthetic systems from the same georeferenced data, enhancing adaptability for addressing various scenarios.

In totality, our approach boasts a commendable computational efficiency, with a total processing time of 460.4 seconds for all procedures. This streamlined process holds promise for advancing research, testing, and planning in power distribution systems, offering valuable data-driven insights for decision-makers and researchers. Importantly, our focus is not on direct comparisons between our computational time and state-of-the-art methods - and for the best of author's knowledge, previous work do not explicitily mention the computational time of creating synthetic power DS - but rather to show the efficiency of our method in creating realistic synthetic distribution systems in a matter of seconds is a notable strength. The scalability of our proposed method is also evident, as demonstrated by its completion within 460 seconds for a substantial test case using georeferenced data from the largest Brazilian city.

TABLE IV
COMPUTATIONAL TIME FOR EACH STEP OF THE PROPOSED APPROACH, AS DELINEATED IN FIGURE 1. THE PROCESS IS ITERATED 100 TIMES, AND THE MEAN COMPUTATIONAL TIME IS COMPUTED FOR ROBUST EVALUATION.

Procedure	Computational Time (seconds)
Procedure 1: Breadth First Search to allocate feeders	125.4
Procedure 2: Allocate normally opened switches	108.2
Procedure 3: Calculate Posterior Distributions	216.0
Procedure 4: Sample from posterior distribution	10.8
Total computational time	**460.4**

V. CONCLUSIONS

The increasing complexity of distribution systems, fueled by factors such as distributed generation, renewable energy integration, and increasing demand, underscores the need for innovative approaches to generate robust synthetic data for power system analysis and control. This study introduces Bayesian Hierarchical Modeling (BHM), emphasizing its strength not just in delivering high-quality results but, critically, in the robustness of the Bayesian methodology. Our approach stands out in the literature by harnessing the methodological power of BHM, treating each parameter as an independent random variable with dedicated distribution curves and hyperparameters. This strategic choice ensures a nuanced understanding of intricacies in power distribution systems, enhancing the generation of synthetic data that faithfully mirrors real-world complexities. The results demonstrate the proposed model's ability to learn from real data, generate posterior distributions for various system characteristics modeled as random variables, and sample from these distributions to create a georeferenced synthetic grid. This highlights the efficacy of our approach in crafting realistic synthetic distribution systems, making a valuable contribution to the advancement of power system analysis and control methodologies.

The BHM's distinctive treatment of parameters, informed by expert knowledge or real-world data, results in posterior distributions intricately linked to their respective highest density intervals (HDIs). The integration of HDI-backed distributions enhances the versatility of synthetic distribution generation and furnishes decision-makers with invaluable insights for well-informed choices. This proves pivotal in guiding investments and system planning, especially in georeferenced urban areas. The outcomes of our model underscore its generalizability and adaptability, driven by two distinct features. Firstly, the incorporation of urban data from an open-source platform, such as OpenStreetMaps, imparts a versatility that extends beyond the confines of any specific city. This flexibility allows the generation of synthetic networks for a diverse array of urban environments, eliminating geographic limitations associated with the choice of São Paulo as a reference city. Secondly, the Bayesian Hierarchical Model (BHM) employed in our approach contributes to its generalization. The BHM exhibits a unique capacity to assimilate new data from any distribution system, dynamically adapting to varying characteristics and complexities. Moreover, the BHM's flexibility allows for modeling based on expert knowledge. By fixing specific values for the parameters of probability distributions, the model can be tailored to incorporate the expert's insights, offering an intuitive and personalized approach to synthetic network generation. Thus, the generalist nature of our model highlights its applicability across a wide range of scenarios, making it a valuable tool for the analysis of energy distribution systems in diverse contexts.

Moreover, a noteworthy aspect of our methodology is its computational efficiency, completing the entire process, including georeferenced data extraction, BHM learning, and posterior distribution sampling, in a mere 460.4 seconds. Once the georeferenced data and the fitted BHM are established, generating new synthetic power systems requires less than 15 seconds. While computational time is not the primary focus of our work, this rapid execution facilitates iterative runs for optimization, contributing to the efficiency of research, testing, and planning in the power distribution domain.

Finally, our synthetic power distribution system generation method has limitations. Its accuracy is sensitive to input data quality, and discrepancies may impact realism. The BHM introduces uncertainties based on expert knowledge and real-world data. Future research directions could enhance the methodology with advanced statistical models, incorporate machine learning for improved data processing, and address scalability issues. Extending the application to diverse urban settings and datasets would bolster generalizability. Additionally, exploring how the proposed BHM can exploit the escalating complexity associated with distributed generation and renewable energy represents a noteworthy avenue for future investigation. This consideration is vital for the ongoing evolution and broader applicability of our approach.

ACKNOWLEDGMENTS

This work was partially supported by: São Paulo Research Foundation (FAPESP), grants 2023/07634-7 and 2021/12220-1; CAPES - Brazil (88887.682748/2022-00). The authors of this work would like to thank the Center for Artificial Intelligence (C4AI-USP) and the support from the IBM Corporation.

REFERENCES

[1] H. K. Cakmak, L. Janecke, M. Weber, and V. Hagenmeyer, "An optimization-based approach for automated generation of residential

low-voltage grid models using open data and open source software," in *2022 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*. IEEE, Oct. 2022.

[2] L. Wang, J. Halvorsen, S. Pannala, A. Srivastava, A. H. Gebremedhin, and N. N. Schulz, "CP-SyNet: A tool for generating customised cyber-power synthetic network for distribution systems with distributed energy resources," *IET Smart Grid*, vol. 5, no. 6, pp. 463–477, Oct. 2022.

[3] OpenStreetMap contributors, "Planet dump retrieved from https://planet.osm.org ," https://www.openstreetmap.org , 2017.

[4] M. Mishra, J. Martinsson, M. Rantatalo, and K. Goebel, "Bayesian hierarchical model-based prognostics for lithium-ion batteries," *Reliability Engineering &amp System Safety*, vol. 172, pp. 25–35, Apr. 2018.

[5] R. Gaugl, S. Wogrin, U. Bachhiesl, and L. Frauenlob, "GridTool: An open-source tool to convert electricity grid data," *SoftwareX*, vol. 21, p. 101314, Feb. 2023.

[6] D. Sarajlic and C. Rehtanz, "Low voltage benchmark distribution network models based on publicly available data," in *2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe)*. IEEE, Sep. 2019.

[7] M. Ali, K. Prakash, C. Macana, M. Raza, A. Bashir, and H. Pota, "Modeling synthetic power distribution network and datasets with industrial validation," *Journal of Industrial Information Integration*, vol. 31, p. 100407, Feb. 2023.

[8] A. Bidel, T. Schelo, and T. Hamacher, "Synthetic distribution grid generation based on high resolution spatial data," in *2021 IEEE International Conference on Environment and Electrical Engineering and 2021 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I&ampCPS Europe)*. IEEE, Sep. 2021.

[9] H. Li, J. L. Wert, A. B. Birchfield, T. J. Overbye, T. G. S. Roman, C. M. Domingo, F. E. P. Marcos, P. D. Martinez, T. Elgindy, and B. Palmintier, "Building highly detailed synthetic electric grid data sets for combined transmission and distribution systems," *IEEE Open Access Journal of Power and Energy*, vol. 7, pp. 478–488, 2020.

[10] E. Schweitzer, A. Scaglione, A. Monti, and G. A. Pagani, "Automated generation algorithm for synthetic medium voltage radial distribution systems," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 7, no. 2, pp. 271–284, Jun. 2017.

[11] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck, "Probabilistic programming in python using PyMC3," *PeerJ Computer Science*, vol. 2, p. e55, apr 2016. [Online]. Available: https://doi.org/10.7717/peerj-cs.55

[12] L. Thurner, A. Scheidler, F. Schäfer, J. Menke, J. Dollichon, F. Meier, S. Meinecke, and M. Braun, "pandapower — an open-source python tool for convenient modeling, analysis, and optimization of electric power systems," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 6510–6521, Nov 2018.

[13] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring network structure, dynamics, and function using networkx," in *Proceedings of the 7th Python in Science Conference*, G. Varoquaux, T. Vaught, and J. Millman, Eds., Pasadena, CA USA, 2008, pp. 11 – 15.

[14] D. L. Marino, C. S. Wickramasinghe, C. Rieger, and M. Manic, "Data-driven stochastic anomaly detection on smart-grid communications using mixture poisson distributions," in *IECON 2019-45th Annual Conference of the IEEE Industrial Electronics Society*, vol. 1. IEEE, 2019, pp. 5855–5861.

[15] E. Schweitzer, A. Scaglione, A. Monti, and G. A. Pagani, "Automated generation algorithm for synthetic medium voltage radial distribution systems," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 7, no. 2, pp. 271–284, 2017.

[16] M. Fogliatto, H. Caetano, L. D. N., J. Massignan, R. Fanucchi, J. London, B. Pereira, M. Bessani, and C. Maciel, "Power distribution system interruption duration model using reliability analysis regression," *Electric Power Systems Research*, vol. 211, p. 108193, Oct. 2022. [Online]. Available: https://doi.org/10.1016/j.epsr.2022.108193