Analyzing Impact of Data Uncertainty in Distributed Energy Resources using Bayesian Networks

Ankita Garg
Department of Computer Science
Durham University, UK
ankita.garg@durham.ac.uk

Gagangeet Singh Aujla
Department of Computer Science
Durham University, UK
gagangeet.s.aujla@durham.ac.uk

Hongjian Sun
Department of Engineering
Durham University, UK
hongjian.sun@durham.ac.uk

***Abstract*—With the high penetration of distributed energy resources (DERs), distribution networks have become more prone to uncertainties associated with renewable energy sources (RESs). If not handled judiciously, these uncertainties may lead to interruption in power supply and even failure of the entire power system in the long run. In this paper, a Bayesian Network (BN) approach is used to find the hidden inter-dependencies among the various weather parameters and how these affect renewable energy generation. A heuristic algorithm is then proposed to identify the root-cause of the uncertainty which increases the overall grid dependency and thereby, tackling its associated carbon emissions. To check the efficacy of the proposed approach, the effect of data uncertainty in the distribution network with DERs penetration in nine different regions of England is discussed. Furthermore, a case-study of a residential area of Newcastle upon Tyne is discussed in detail to back-trace the root-cause of the fault that occurred in one of the DERs.**

***Index Terms*—Bayesian network, root-cause analysis, DERs, data uncertainty, carbon emissions, energy bills**

I. Introduction

Now-a-days, distributed energy resources (DERs) are gaining importance due to their affordability and reliability to meet the energy demand. DERs facilitate the use of energy efficiently by generating it on-site and storing it for the use during peak-operating times [1], [2]. However, the inherent uncertainties involved with renewable energy sources (RESs), such as power fluctuations and intermittence, may deteriorate the stability and security of power grids [3]. Various techniques to model these uncertainties in energy systems including the probabilistic method, fuzzy variable approach, hybrid optimization, and non-linear programming are discussed in [3]–[5]. Nevertheless, increased deployment of DERs may still result in congestion within the distribution grid, which would require their active management and control. This situation necessitates a reevaluation of the role of distribution system operators (DSOs) who have traditionally been responsible for planning, maintaining, and managing networks, including handling of outages [6], [7].

DSOs have the potential to act as a central platform for overseeing consumer data concerning electricity usage, generation, billing, location, and the types of DERs involved. By adhering to regulatory standards, DSOs can effectively collect and store this data while upholding consumer rights, including privacy protection [7]. By utilizing this data, DSOs can enhance their ability to forecast demand, resulting in improved planning and operation of the system. This data can facilitate increased adoption of RESs by empowering consumers to gain insights into their energy consumption and/or production patterns.

Furthermore, as the optimal and reliable operation of DERs based distribution network has become dependent on the data inflows, any fiddling with the data has the capability to make the entire energy network unstable. Therefore, it is crucial that DSOs get the correct data information at each time step. In [8], the concept of geosensors was introduced to model and predict the sensor data uncertainty for the environmental monitoring applications. In [9], authors proposed the guide to the expression of measurement uncertainty and Adaptive Monte Carlo method, to analyze the measurement uncertainty in underwater positioning systems. However, the uncertainty evaluation in DERs requires higher data granularity to identify the real-time data variations, in which the existing methods would fail [3]–[5], [8], [9].

Bayesian Networks (BNs) are being used in various applications due to their ability to provide reliability in uncertainty estimations. Many researchers have used BNs in the energy sector for various applications including energy forecasting, optimization risk assessment, O&M planning, etc. [10], [11]. However, modeling the data uncertainty using BNs in DERs is still an unexplored research area. Motivated by this, in this paper, we would be focusing on the impact of data uncertainty in the DERs that could affect the operation of the distribution network. The major contributions of this paper are:

- Bayesian Network based approach is used to analyze the correlation among various weather parameters and fit their probability distribution using metalog distribution.
- Root-cause of increasing grid dependency and carbon emissions is analyzed using a heuristic approach.

The rest of the paper is organized as follows. Section II presents the system modeling and Section III describes the proposed methodology. Results are discussed in Section IV while the paper is concluded in Section V.

II. System Modeling

A. Power system model with DER Integration

With the emergence of DERs such as rooftop solar panels, battery storage, electric vehicles, and wind turbines, the end-users are participating in the energy market by selling and

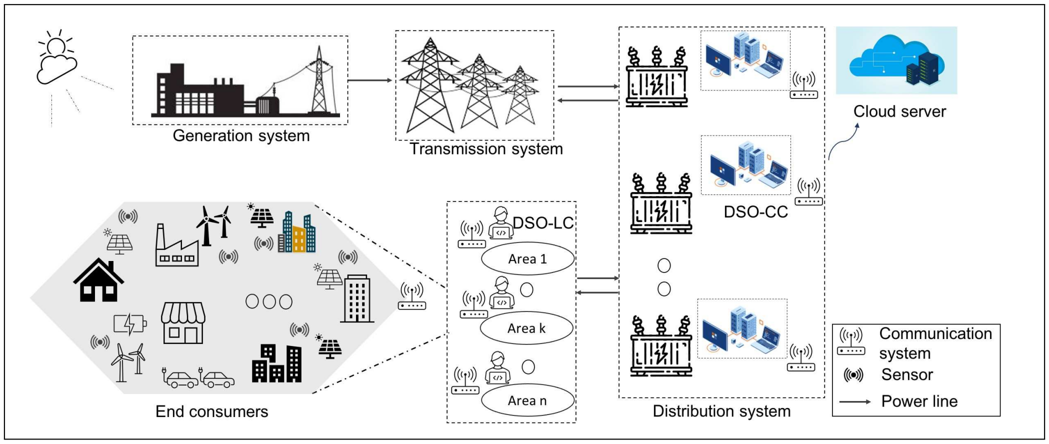


Fig. 1: Power system model with DERs integrated distribution network

purchasing energy from the power grid, allowing two-way flow of power. Figure 1 shows the power system that has a power generating station, and the power is transmitted via the transmission network to the distribution network. The distribution network is integrated with DERs and is connected with the end-users with DERs. For the sake of simplicity, we have considered only four areas (with DERs) in each of the considered nine regions of England. Each area is equipped with DERs to meet its load demand. DSO-Local controller (LC) is responsible to manage and control the incoming requests from all the areas. DSO-central controller (CC) continuously run checks on all the data sent by all the LCs and saves that in the cloud server to compute the day-ahead energy scheduling for the entire distribution network.

The solar power ($P_{pv}(t)$) of each solar panel of an area at time t is calculated as [12]:

$$P_{pv}(t) = \beta P_{peak} D_{pv} \left[\frac{G_T(t)}{G_{STC}} \right] [1 - \alpha_p (T_C - T_{C,STC})] \quad (1)$$

where P_{peak} is the peak PV module output power, D_{pv} is the derating factor (%), α_p is the power temperature coefficient, $G_T(t)$ and G_{STC} are solar radiance incident on PV module at time t (kW/m^2) and under standard testing conditions respectively, T_C and $T_{C,STC}$ are PV panel temperature and PV cell temperature under standard testing conditions respectively, $\beta = N_s \times N_p$, i.e., number of panels connected in series and parallel respectively.

The real-time wind power (P_w) of each wind turbine of an area at time t is expressed as [13]:

$$P_w(t) = \begin{cases} 0 & ; \quad V(t) \leq V_i \text{ or } V(t) \geq V_o \\ \gamma P_w^r (\frac{V(t) - V_i}{V_r - V_i}) & ; \quad V_i \leq V(t) \leq V_r \\ \gamma P_w^r & ; \quad V_r \leq V(t) \leq V_o \end{cases} \quad (2)$$

where $V(t)$ is wind speed at time t, V_i & V_o are cut-in & cut-off wind speeds, V_r is rated wind speed, and P_w^r is rated wind turbine power. γ is the number of wind turbines installed.

B. Data Model

To simulate the energy model, data from the selected areas of England is gathered and then processed as follows:

1) Site selection: In this paper, to illustrate the impact of data uncertainty, the map of England is divided into nine different geographical locations (based on distinct weather conditions) as shown in Fig. 2. The locations are selected prudently to analyze and simulate the effect of uncertain and complex weather conditions on England's energy network.

Fig. 2: Nine regions in the map of England

2) Data Collection: The data is collected for 1 year (15-May-2022 to 15-May-2023) [14] for nine locations in CSV format. Each location data is saved in a different file which has 14 different weather parameters. Furthermore, the load demand data of each location is generated synthetically using HOMER software. It is assumed that the sensors installed at each DER are giving the generation data estimated using Eq. (1) and Eq. (2) (solar and wind).

3) Data discretization: The data is then discretized based on a hierarchical method to convert the huge amount of data into finite states to increase its intractability. Discretization makes it easy to interpret real-world examples without having expert knowledge. For instance, solar irradiance is discretized into three states, i.e., low, medium, and high, which gives a clear idea to the reader to interpret the results.

III. Proposed Methodology

In this section, the basics of BNs to estimate the conditional probabilities of random (uncertain) variables is discussed. Further, the PDFs of these uncertain variables are estimated using metalog distribution, to estimate their probabilities that are used to find out the correlations among different uncertain variables using Tree-augmented naive Bayes.

A. Fundamentals of Bayesian Networks

BNs are directed acyclic graphs (DAGs) that generate correlations amongst random variables in the form of nodes and links. Bayes Theorem forms the basis of BNs, which can update the probability of a random variable based on its existing evidence [10]. Graphically, these scenarios are represented in the form of directed graphs, wherein uncertain parameters form the nodes and the edges describe the correlation between two nodes. Mathematically, according to Bayes theorem, posterior probability or the conditional probability of occurrence of an event A given the probability of event B ($Pr(B)$) is [11]:

$$Pr(A|B) = \frac{Pr(B|A).Pr(A)}{Pr(B)} \tag{3}$$

where, $Pr(B|A)$ is the conditional probability of event B given the probability of an event A ($Pr(A)$).

B. Fitting Metalog Distributions using Bayesian Networks

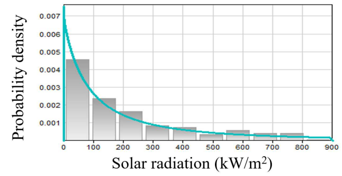
Some commonly used continuous probability distributions include the normal, lognormal, Weibull, gamma, and beta distributions, (often referred to as named distributions) are employed to represent univariate data [15]. An expert knowledge is required to select the PDF to fit a data set generated from an unknown or uncertain process. Therefore, in this paper, we have used metalog distributions to fit the data. The metalog distribution offers shape-flexibility, i.e., it best matches the data from any unknown source as compared to the traditional distributions [15]. The following steps are followed to fit the metalog distribution of the considered dataset:

- Determine the upper and lower bounds of the parameter whose distribution is to be fitted.
- Select the number of terms (k), that determines the shape of the distribution. If $k< 6$, the PDF would be smooth as compared to larger values of k, which could fit the data with multiple peaks.
- Finally, the data is fitted according to the metalog quantile function.

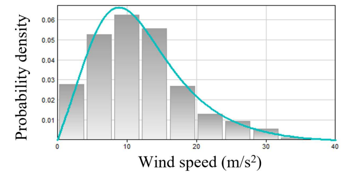
The metalog distributions of solar irradiance and wind speed with $k = 2$ are shown in Fig. 3a and Fig. 3b respectively.

C. Tree Augmented Naive Bayes Algorithm

Once the PDFs of uncertain variables are estimated, their conditional probabilities are calculated to find the inter-dependencies among them using the Tree-Augmented Naive Bayes algorithm, which is a form of semi-naive Bayesian Learning. In this method, each node is associated with the class (parent node) and among each other, allowing correlations among the nodes to be captured [16]. Figure 4 shows an implementation of the Tree-augmented naive Bayes algorithm, in which each weather parameter is directly correlated with solar radiation (class variable) and other parameters. The correlation among the different nodes is essential in this study, as the nodes with similar relationships with solar radiation or wind speed can be eliminated, to simplify the system calculations.



(a) solar radiation



(b) wind speed

Fig. 3: Metalog probability distribution

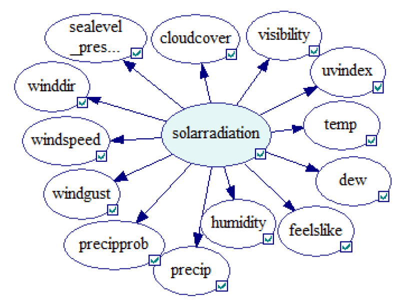


Fig. 4: Dependency graph of weather parameters

D. Root cause analysis using Bayesian Networks

A heuristic algorithm (algorithm 1) is proposed in this paper to determine the root cause of the increase in energy bills and carbon emissions at the distribution end. The problem is framed as:

- **Effect**: Unexpected incoming request at LC
- **Consequence**: Imbalance in expected energy delivered.
- **Cause**: Uncertainty detected at LC end

Local Controller (LC) at the regional level checks the real-time power sold (P_{sold}^{rt}) and purchased (P_{pur}^{rt}) from the grid of each of its areas. When LC encounters a mismatch between the real-time and expected power values, it runs a check on real-time power flows in each area. The real-time powers generated from solar and wind (can be calculated using Eq. (1) and Eq. (2) respectively) are compared with expected solar ($P_{pv}^{exp}(t)$) and wind ($P_{w}^{exp}(t)$) power. The expected values are the forecasted values obtained from the optimal energy scheduling of the DERs at time t.

If the power generation values match, then the expected and real-time power demand is checked. If both matches, then LC sends an okay signal to the CC (central controller), stating that there is no issue at the local server. However, if the expected power demand ((P_{load}^{exp})) does not match with the real-time power demand (P_{load}^{rt}), then this scenario comes under behavioral uncertainty, i.e., there is an unexpected change in the load demand based on the occupant's behavior.

Algorithm 1 Root-cause analysis of uncertainty using BNs

Input: Weather parameters, real-time data, energy scheduling data
Output: Uncertainty detected

Check power purchased and sold to the grid at each time-step for all areas $A_i^t \in i = 1, ..., n$ and $t = 1, ..., h$
while $\forall\ A_i^t$ **do**
  if ($P_{pur}^{exp}(t)$ = $P_{pur}^{rt}(t)$) **then** ▷ *Check total power*
    Run normal operation
  else if ($P_{solar}^{exp}(t) \neq P_{solar}^{rt}(t)$) **then**
    for $\forall DER_j$ **do** ▷ *Check solar for j=1 to M*
      Calculate conditional probabilities of weather parameters using Eq. (3) and compute (G_{min}, G_{max})
      Compute $P_{pv}(t)$ for (G_{min}, G_{max}) using Eq. (1) and return $(P_{pv}^{min}(t), P_{pv}^{max}(t))$
      if ($P_{pv}^{rt} \notin (P_{pv}^{min}(t), P_{pv}^{max}(t))$) **then**
        Data uncertainty in DER_j
      end if
    end for
  else if ($P_{wind}^{exp}(t) \neq P_{wind}^{rt}(t)$) **then**
    for $\forall DER_j$ **do** ▷ *Check wind for j=1 to N*
      Calculate conditional probabilities of weather parameters using Eq. (3) and compute (V_{min}, V_{max})
      Compute $P_w(t)$ for (V_{min}, V_{max}) using Eq. (2) and return $(P_w^{min}(t), P_w^{max}(t))$
      if ($P_w^{rt} \notin (P_w^{min}(t), P_w^{max}(t))$ **then**
        Data uncertainty in DER_j
      end if
    end for
  else if ($P_{load}^{exp}(t) \neq P_{load}^{rt}(t)$) **then**
    Behavioural uncertainty
  else
    Send an okay signal to the central controller
  end if
end while

In contrast, if either real-time solar or wind power generation does not match with their respective expected value (with a pre-determined tolerance limit), then LC would run the Bayesian analysis. The Bayesian analysis would give a probability of solar or wind power generation to be in a certain range. If the real-time values lie under that range, then there are high chances of weather uncertainty. Otherwise, the mismatch is due to faulty data. The reasons for faulty data are:

- **Faulty sensor:** When the sensor of an area gives random or same values for every time step.
- **Dead Sensor:** When the sensor is giving zero value, even if there is some expected value.
- **Data Manipulation:** When someone deliberately manipulates the data to make money.

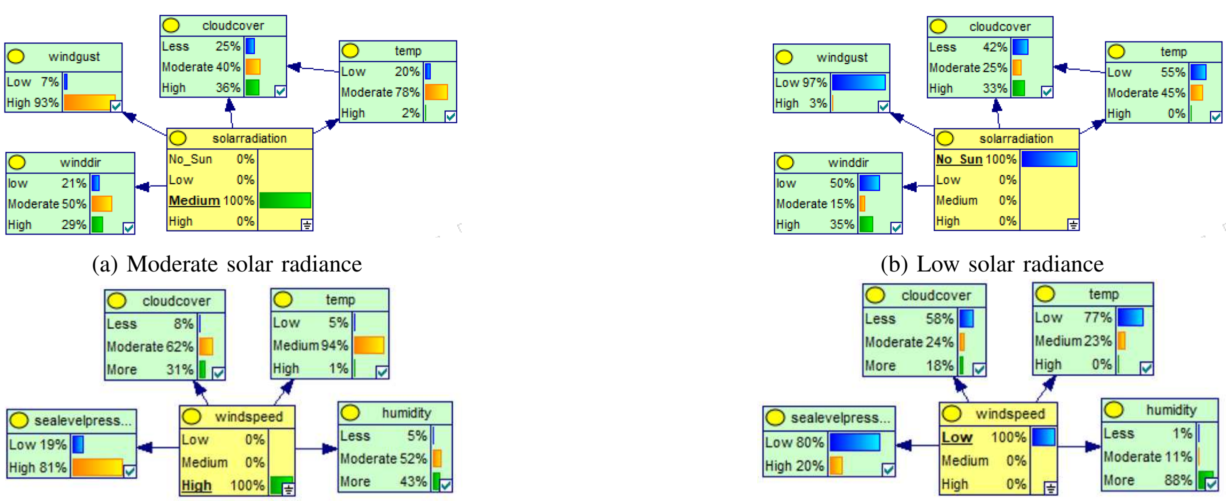
IV. Results and Discussion

The BN analysis was carried out individually for the 4 areas in the nine regions in England. In this paper, the simulation results for one area in Newcastle upon Tyne have been discussed in detail and the same methodology can be applied to determine the erroneous data for the other regions. The results obtained were categorized into 3 phases which are discussed as follows:

A. Phase A: Correlation among different weather parameters

In this phase, the simulation was run to find the inter-dependency graph for different weather parameters (as described in section III). Furthermore, assigning solar radiation as the decision node, its correlation with other parameters was analyzed under two conditions, i.e., 1) low solar radiation, and 2) high solar radiation. Out of the 13 other weather parameters, only 4 parameters (cloud cover, wind gust, temperature, and wind direction) showed major variations with the change in solar radiation. Similarly, with wind speed as the decision variable, the 4 parameters that showed high correlation were sea level pressure, cloud cover, temperature, and humidity. The inferences from this phase were:

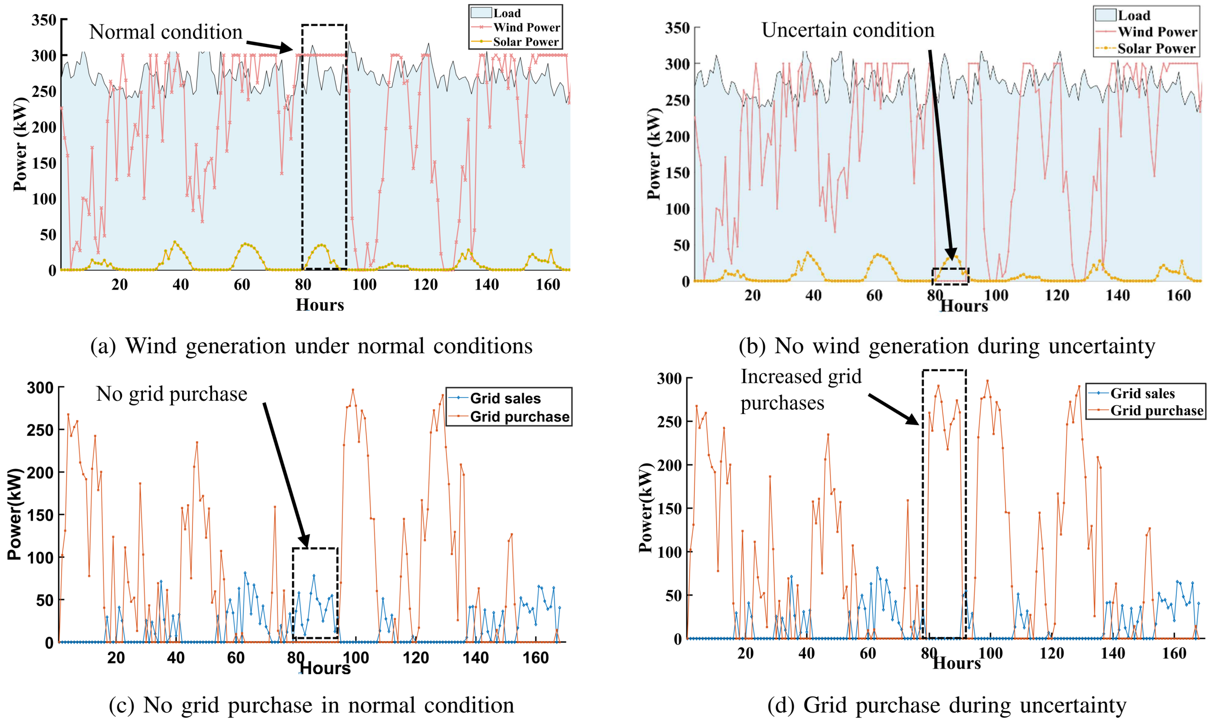
- Fig. 5b depicts that when solar radiation was low, the probability that wind gust, cloud cover, temperature, and wind direction were low was 97%, 42%, 55%, and 50% respectively. Whereas when the solar radiation was in the medium range (as shown in Fig. 5a); cloud cover, temperature, and wind direction changed to the moderate range



(a) Moderate solar radiance (b) Low solar radiance

(c) High wind speed (d) Low wind speed

Fig. 5: Bayesian analysis results showing correlation of weather parameters



(a) Wind generation under normal conditions

(b) No wind generation during uncertainty

(c) No grid purchase in normal condition

(d) Grid purchase during uncertainty

Fig. 6: Energy profiles under different conditions

with the probabilities of 40%, 78%, and 50% respectively, and chances of wind gust in the high value was 93%.
- There were 88%, 58%, and 77% chances that sea level pressure, cloud cover, and temperature respectively were low when wind speed was low (as depicted in Fig. 5d). However, humidity was high with a probability of 88%. In contrast in Fig. 5c, when wind speed was high, the values of cloud cover, temperature, and humidity were in the moderate range with the probabilities of 62%, 94%, and 52%. The sea level pressure was high with a probability of 81% when the wind speed was high.

B. Phase B: Power generation under data uncertainty

The effect of uncertainty on the power flows (i.e., renewable energy generation, energy sold and purchased from the grid) in one of the residential complexes in Newcastle upon Tyne, was determined in this phase. Erroneous data was synthetically injected (wind speed from $t = 80$ h to $t = 90$ h is made zero) in the energy model to analyze and validate the impact of data uncertainty at the consumer end and how it could propagate to affect the entire distribution network. A PSO-based optimization algorithm [17] was used to simulate the system parameters and to calculate the renewable energy generation, load demand, grid sales, and purchase at each time step (equals to 1 hour). The inferences from this phase were:

- Fig. 6 shows the solar and wind generation for 7 days and the corresponding energy sold to the grid and purchased from the grid, under normal conditions, i.e., when no uncertainty was injected in the system. During $t = 80$ h to $t = 90$ h, wind energy generation itself was sufficient to meet the load demand and the surplus energy was sold to the grid.
- When wind energy generation was zero ($t = 80$ h to $t = 90$ h) as shown in Fig. 6b, there was no sufficient renewable generation and the area gets dependent on grid to meet its load demand. Fig. 6d depicts an increase in grid purchases during uncertainty as compared to the normal case.

C. Phase C: Root-cause determination

The results from the above two phases were combined to find out the DER in a particular area with the faulty sensor contributing to data uncertainty. From algorithm 1 discussed in section III, when there was an energy purchase request from area 1 of Newcastle, LC would check the energy scheduling data at that time interval and perform the following tasks:

- Using the results in Phase B, LC analyzed that there must be energy sold to the grid rather than purchased from the grid under normal conditions.
- Next, it checks the Bayesian results discussed in phase A, which showed the inter-dependencies among the weather parameters, based on which the wind speed (during $t = 80$ h to $t = 90$ h) must be in the high range.
- Using Eq. (2), the wind energy generation was calculated for each DER based on the current wind speed value, which was in the range as per the Bayesian equation.
- LC then finds the DER that is giving zero value due to the damaged sensor.

To validate the precision of the proposed method, 10% of the data was altered to inject different types of data uncertainty across 9 cities as shown in Table I. In the case of a dead sensor, these data values were made zero, and the simulations were reciprocated for repeatability of results. It was observed that the proposed method was able to detect the dead sensor uncertainty in all the cases. Further, the method successfully identified the error due to a faulty sensor (in which the data was giving the same value). The method worked well in cases when the injected data values were increased by 20% and 50% respectively, however, failed to identify the fault when increased by just 5%. This failure might be due to the reason that BNs work with a range of values, and 5% increase in the data was within the tolerance limit of the method.

TABLE I: Evidence table for proposed method

Place	**Type of erroneous data**	**Injected uncertainty**	**Detected**
Newcastle	Dead sensor	Zero value	✓
York	Faulty sensor	Same value	✓
Manchester	Data Manipulation	20% ↑ in value	✓
Birmingham	Dead sensor	Zero value	✓
Norwich	Faulty sensor	Same value	✓
Bristol	Data Manipulation	5% ↑ in value	×
London	Dead sensor	Zero value	✓
Southampton	Faulty sensor	Same value	✓
Plymouth	Data Manipulation	50% ↑ in value	✓

Overall, the results in this paper highlight the impact of data uncertainty at the local end. If this is not evaluated at the local server, this uncertainty could propagate to the central level. This tiny faulty sensor could impact each level in the distribution network as:

- *Impact on End-User:* As the grid purchase increase, energy bills will increase. During peak demand hours, the data uncertainty could further increase the bills, which results in hefty bills to the consumer.
- *Impact on Distribution level:* Due to power imbalance at the faulty node, energy scheduling needs to be re-calculated, increasing the computational burden on CC.
- *Impact on Environment:* More the dependency on the grid, the more would be the emissions. Hence, failure to scrutinize data uncertainty pollutes the environment and hinders our target to achieve net-zero emissions.

V. Conclusion

The penetration of DERs in the distribution network transforms the conventional distribution grid and enhances its flexibility to integrate intermittent RESs. In this paper, we have discussed the impact of data uncertainty in the DERs installed in the areas of England and how the root-cause analysis can be performed using BNs. BN analysis performed in this paper utilized the inter-dependencies of various weather parameters among each other to perform the root-cause analysis for detecting which DER is sending the erroneous data. By implementing this work in real-time, one could detect faults in smart energy systems due to malfunctioning IoT devices, which would reduce emissions and costs.

In the future, this work can be extended to mitigate data uncertainty in the distribution and transmission network. Moreover, the applicability of BNs to identify the instances of cyber-attacks on the power system would be analyzed.

Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 872172 (TESTBED2 project: www.testbed2.org), Durham Start Up Grant (S200021) and Durham Seedcorn project (RQ090002).

References

[1] Q. Shi, W. Liu, B. Zeng, H. Hui, and F. Li, "Enhancing distribution system resilience against extreme weather events: Concept review, algorithm summary, and future vision," *International Journal of Electrical Power & Energy Systems*, vol. 138, p. 107860, 2022.

[2] E. F. Bødal, P. C. Del Granado, H. Farahmand, M. Korpås, P. Olivella, I. Munné, and P. Lloret, "Challenges in distribution grid with high penetration of renewables," *NTNU Open*, 2018. [Online]. Available: http://hdl.handle.net/11250/2589060

[3] Y. Zhang, J. Wang, and Z. Li, "Uncertainty modeling of distributed energy resources: Techniques and challenges," *Current Sustainable/Renewable Energy Reports*, vol. 6, pp. 42–51, 2019.

[4] S. Prabhakar and R. Cheng, *Data Uncertainty Management in Sensor Networks*. Boston, MA: Springer US, 2009.

[5] P. F. Pelz, P. Groche, M. E. Pfetsch, and M. Schaeffner, *Mastering uncertainty in mechanical engineering*. Springer Nature, 2021.

[6] A. Hentunen, J. Forsströmm, and V. Mukherjee, "Smart system of renewable energy storage based on integrated EVs and batteries to empower mobile, distributed and centralised energy storage in the distribution grid," *Horizon*, 2020.

[7] C. Zhang, J. Li, Y.-J. A. Zhang, and Z. Xu, "Expanding the role of DSOs in establishing flexible local markets: A brokerage perspective," *IEEE Transactions on Power Systems*, pp. 1–14, 2023.

[8] C. C. G. Rodríguez and S. Servigne, "Managing sensor data uncertainty: a data quality approach," *International Journal of Agricultural and Environmental Information Systems*, vol. 4, no. 1, pp. 35–54, 2013.

[9] J. Zhu, V. Korochentsev, B. Chen, H. Li, and Y. Zhang, "Metering method and measurement uncertainty evaluation of underwater positioning system in six degrees of freedom space," *IEEE Access*, vol. 8, pp. 21 703–21 716, 2020.

[10] S. Hosseini and M. Sarder, "Development of a bayesian network model for optimal site selection of electric vehicle charging station," *International Journal of Electrical Power & Energy Systems*, vol. 105, pp. 110–122, 2019.

[11] T. Adedipe, M. Shafiee, and E. Zio, "Bayesian network modelling for the wind energy industry: An overview," *Reliability Engineering & System Safety*, vol. 202, p. 107053, 2020.

[12] N. M. Kumar, S. S. Chopra, A. A. Chand, R. M. Elavarasan, and G. Shafiullah, "Hybrid renewable energy microgrid for a residential community: A techno-economic and environmental perspective in the context of the SDG7," *Sustainability*, vol. 12, no. 10, p. 3944, 2020.

[13] T. Ma and M. S. Javed, "Integrated sizing of hybrid pv-wind-battery system for remote island considering the saturation of each renewable energy resource," *Energy conversion and management*, vol. 182, pp. 178–190, 2019.

[14] "Weather data," https://www.visualcrossing.com/weather/weather-data-services, accessed: 2023-05-03.

[15] A. K. Khamees, A. Y. Abdelaziz, M. R. Eskaros, M. A. Attia, and A. O. Badr, "The mixture of probability distribution functions for wind and photovoltaic power systems using a metaheuristic method," *Processes*, vol. 10, no. 11, p. 2446, 2022.

[16] "Bayesian networks," https://www.bayesfusion.com/genie/, accessed: 2023-04-24.

[17] H. Cai, S. Shen, Q. Lin, X. Li, and H. Xiao, "Predicting the energy consumption of residential buildings for regional electricity supply-side and demand-side management," *IEEE Access*, vol. 7, pp. 30 386–30 397, 2019.