Smart Meter Data Mining for Peak Load Analysis and Outage Detection in Distribution Systems

Zhao-Yu Wang

Harpole-Pentair Assistant Professor

Iowa State University

(wzy@iastate.edu)

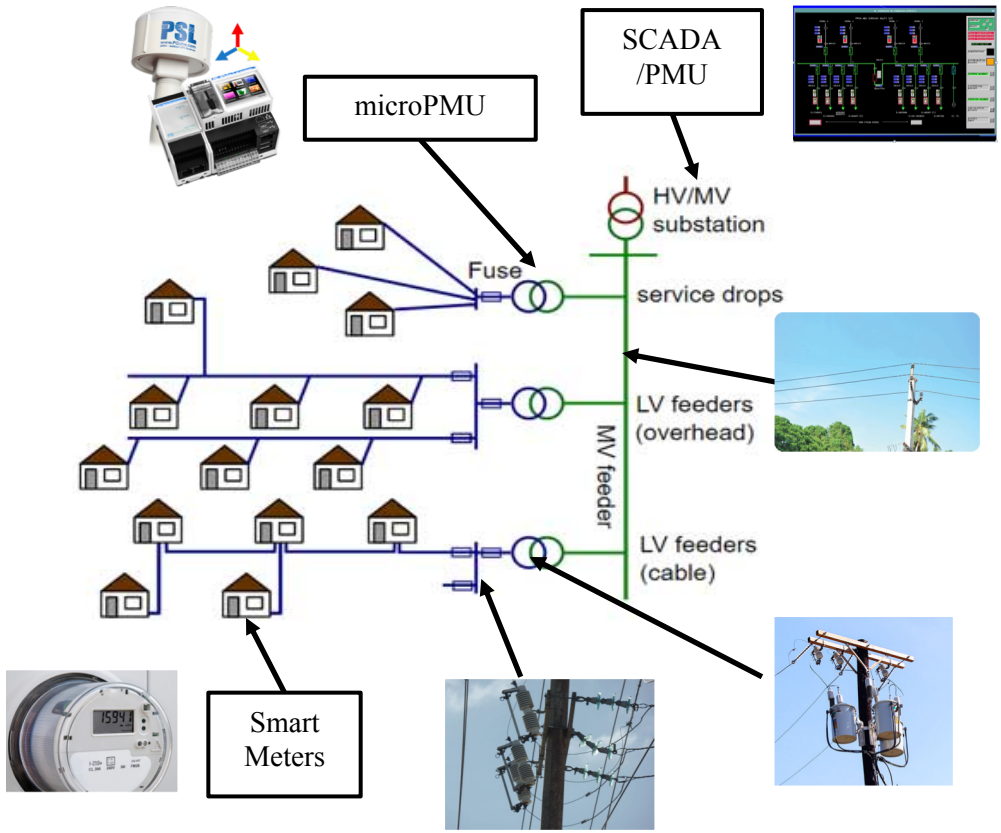
PSERC Webinar

October 13, 2020

Presentation Outline

- **Introduction to Smart Meter Data**
- **Estimating Unobservable Customers' Contributions to System Peak Demand**
- **Outage Detection in Partially Observable Systems**
- **Conclusion and Future Work**

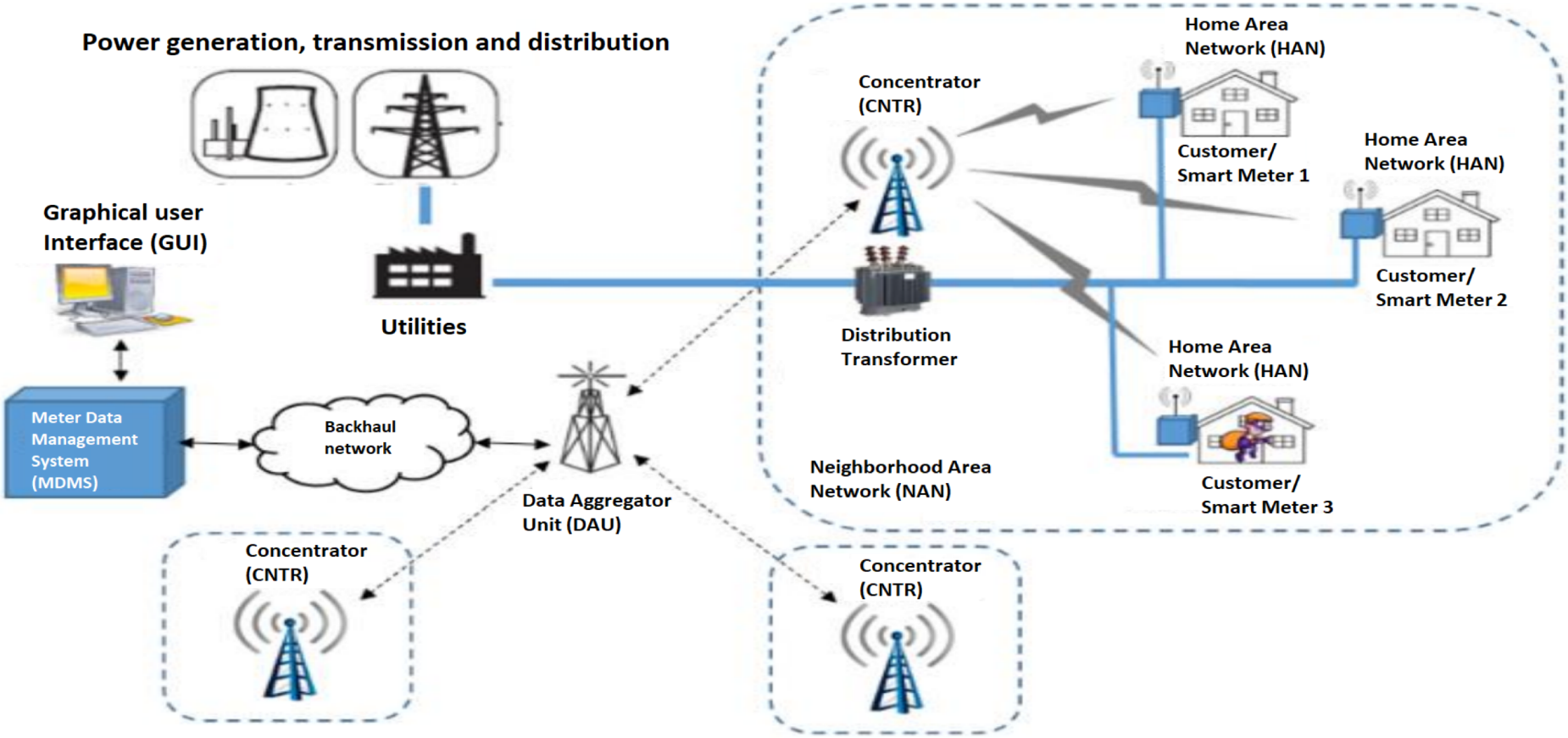
Data in Power Distribution Grids



A Power distribution grid with different sensors

- **Where does the data come from?**
  - SCADA (supervisory control and data acquisition); Smart Meters; Protection Devices; (micro)PMUs (phasor measurement units)
  - Measures voltage/current/frequency at different resolutions
- **What are smart meters?**
  - Stay in your homes
  - Measure energy and voltage, sometimes reactive power
  - 5/15/30/60-minute resolution
  - Single phase or three-phase (large commercial and industrial)

Smart Meter Data Collection



K. K. Kee, S. M. F. Shahab and C. J. Loh, "Design and development of an innovative smart metering system with GUI-based NTL detection platform"

Exemplary Real Data from Utilities

- More AMI data and circuit models:

Utilities	Substations	Feeders	Transformers	Total Customer	Customers with Meters
3	5	27	1726	9118	6631

- Duration: 4 years (2014 - 2018)
- Measurement Type: Smart Meters and SCADA
- **Detailed circuit models of all feeders in Milsoft/OpenDSS and exact smart meter locations**
- Data Time Resolution: 5 Minutes – 1 Hour
- Customer Type:

Residential	Commercial	Industrial	Other
84.67%	14.11%	0.67%	0.55%

Smart Meter Data Pre-Processing

- **Smart Meter Data Problems:**
  - ➢ Outliers/Bad Data
  - ➢ Communication Failure
  - ➢ Missing Data

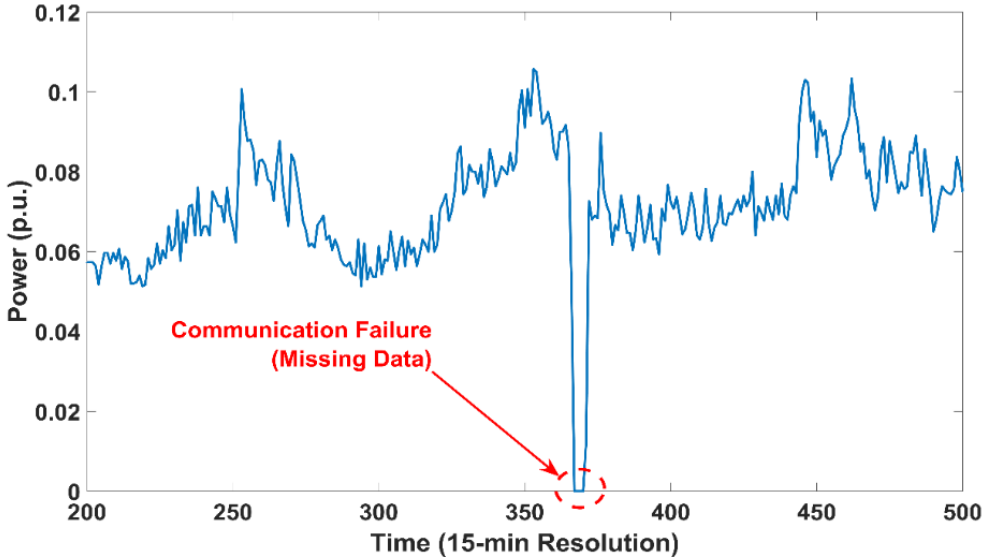
- **Counter-Measures:**
  - ✓ Engineering intuition (data inconsistency)
  - ✓ Conventional Statistical Tools

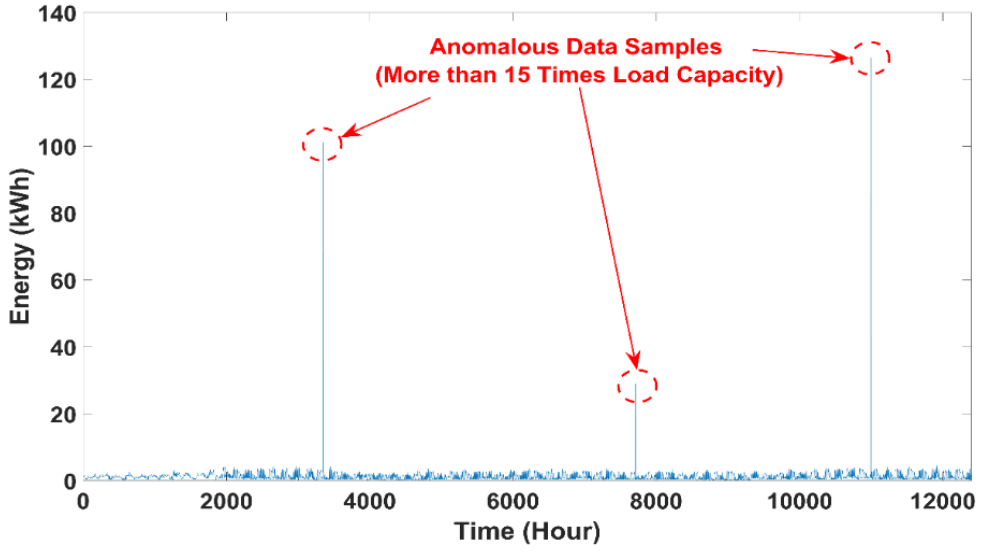
(e.g. Z-score)

  - ✓ Robust Computation

(e.g. relevance vector machines)

  - ✓ Anomaly Detection Algorithms

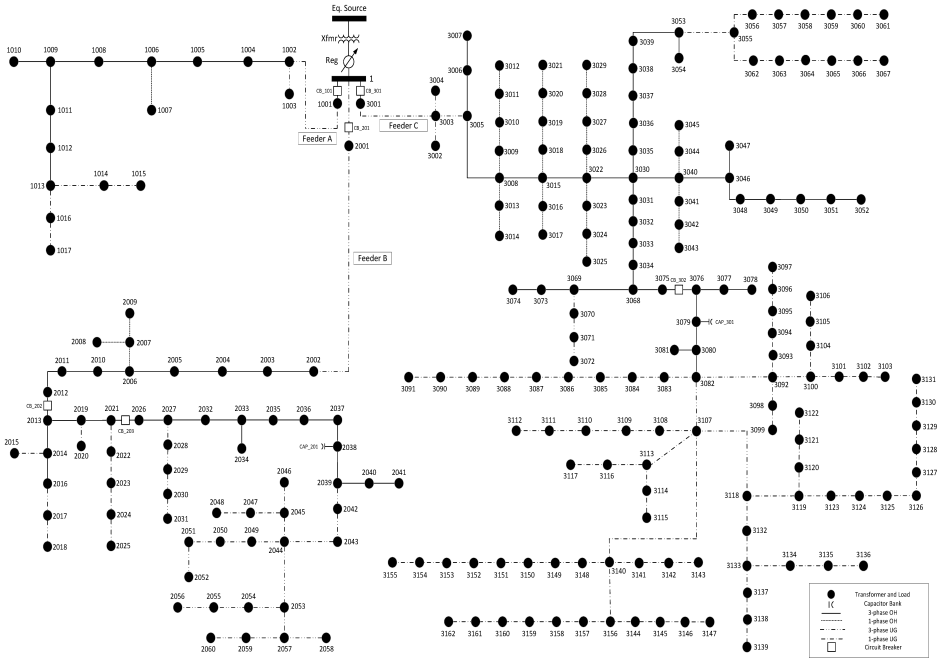




Data Sharing

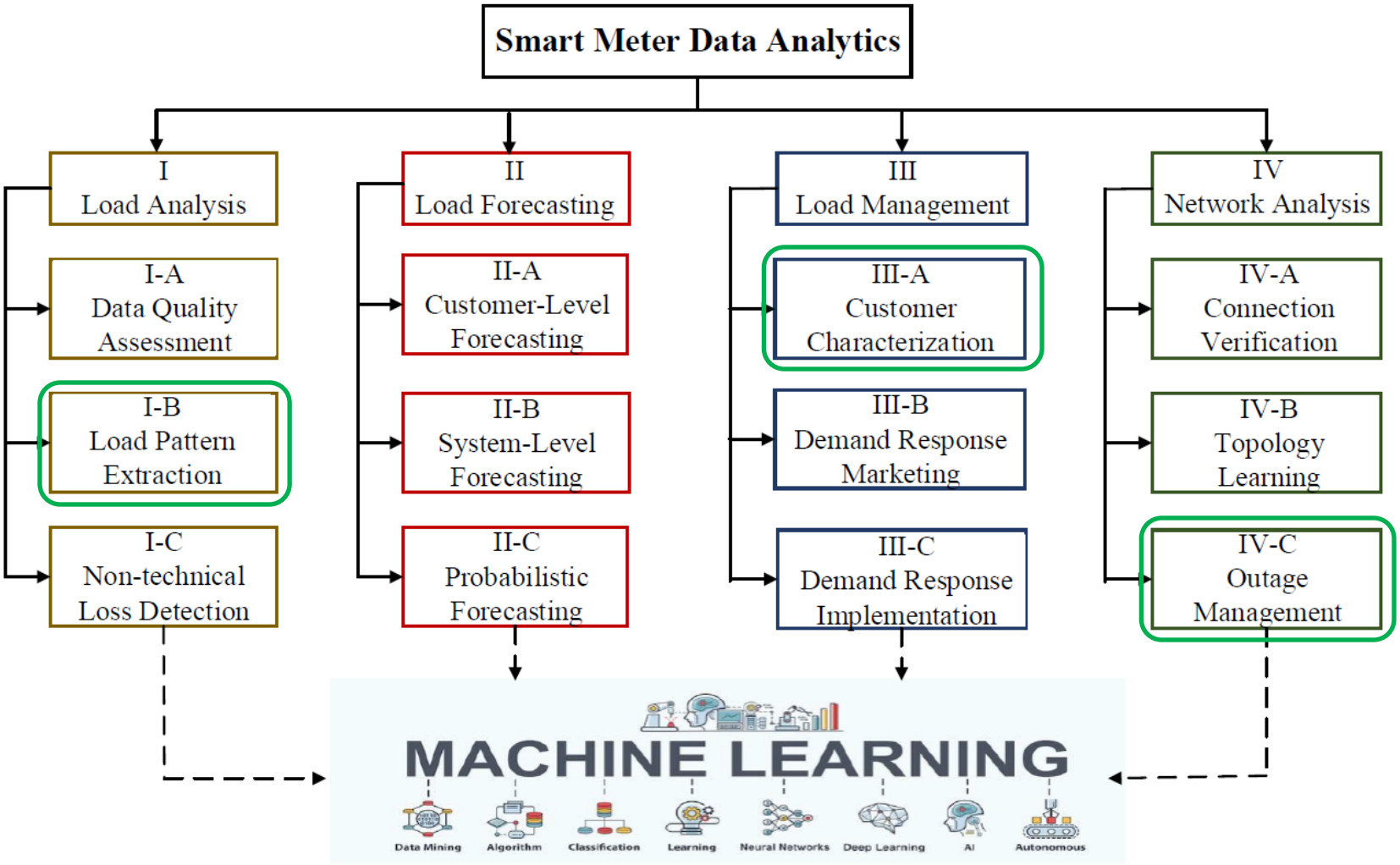
With permission from our utility partner, we share a real distribution grid model with one-year smart meter measurements. This dataset provides an opportunity for researchers and engineers to perform validation and demonstration using real utility grid models and field measurements.

- The system consists of 3 feeders and 240 nodes and is located in Midwest U.S.
- The system has 1120 customers and all of them are equipped with smart meters. These smart meters measure hourly energy consumption (kWh). We share the one-year real smart meter measurements for 2017.
- The system has standard electric components such as overhead lines, underground cables, substation transformers with LTC, line switches, capacitor banks, and secondary distribution transformers. The real system topology and component parameters are included.
- You may download the dataset at: http://wzy.ece.iastate.edu/Testsystem.html , including system description (in .doc and .xlsx), smart meter data (in .xlsx), OpenDSS model, and Matlab code for quasi-static time-series simulation.



Iowa Distribution Test System

What can be learned from smart meter data to improve distribution system operation?



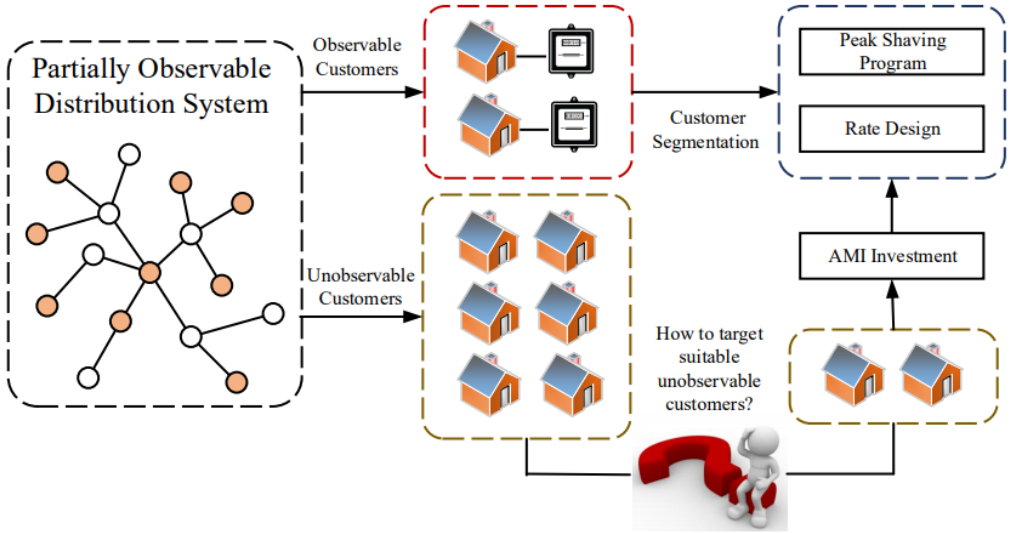
Estimating Customer Peak Contribution

Problem Statement: Inferring *unobservable* residential customers' peak contributions using their monthly energy bills.

Application: Intelligently targeting customers for peak shaving programs, smart meter (SM) investment, and rate design.

Challenges:

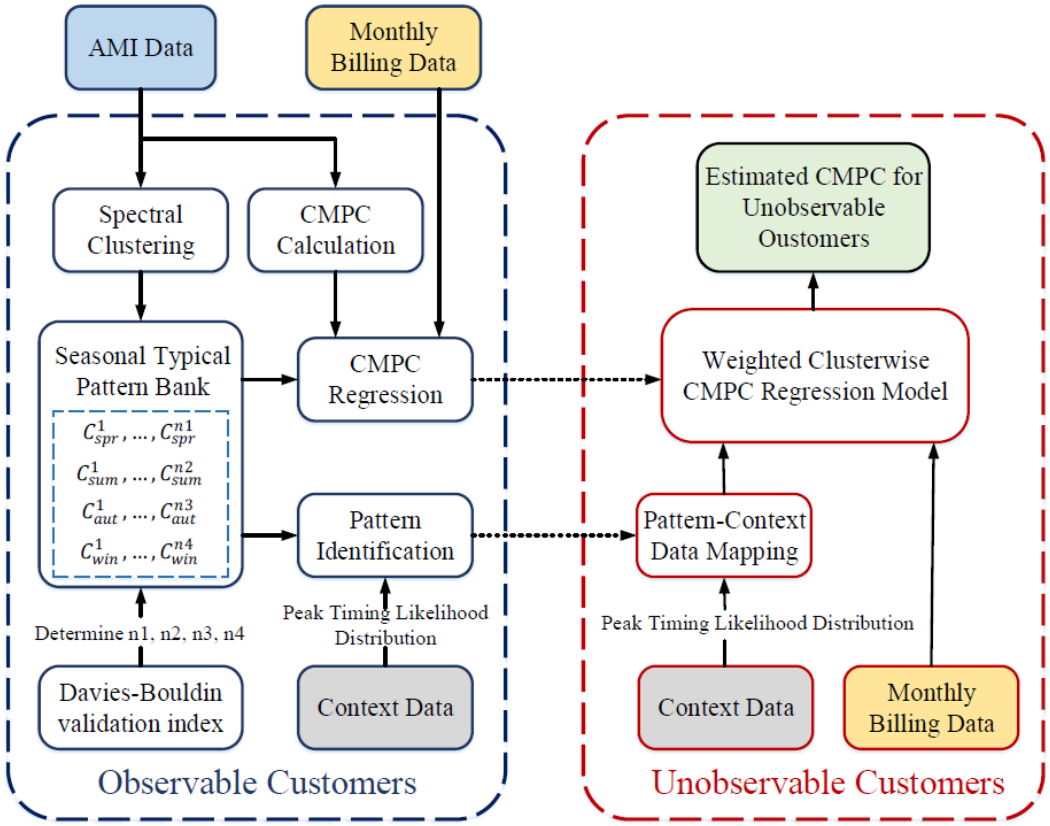
- ✓ System is partially observable – no SM for unobservable customers.
- ✓ Customers with high monthly bills do not necessarily have high peak load.
- ✓ Customers with high peak load do not necessarily have high ***peak contribution*** due to the noncoincidence between customers' and system's peak time.



A Data-Driven Customer Segmentation Strategy Based on Contribution to System Peak Demand

Our Solution:

- Propose a new metric, coincident monthly peak contribution (CMPC), to quantify the customer peak contribution.
- Propose a three-stage method to infer CMPC for unobserved customers using their monthly billing information and context data.
- Context data: sociodemographic data.



CMPC Regression: Mapping CMPC to customer billing data for each typical pattern

Pattern Identification: Mapping context data to typical patterns

Y. Yuan, K. Dehghanpour, F. Bu, and Z. Wang, "A Data-Driven Customer Segmentation Strategy Based on Contribution to System Peak Demand," IEEE Trans. on Power Systems, accepted for publication.

Coincident Monthly Peak Contribution (CMPC)

Customer peak contribution index, CMPC:

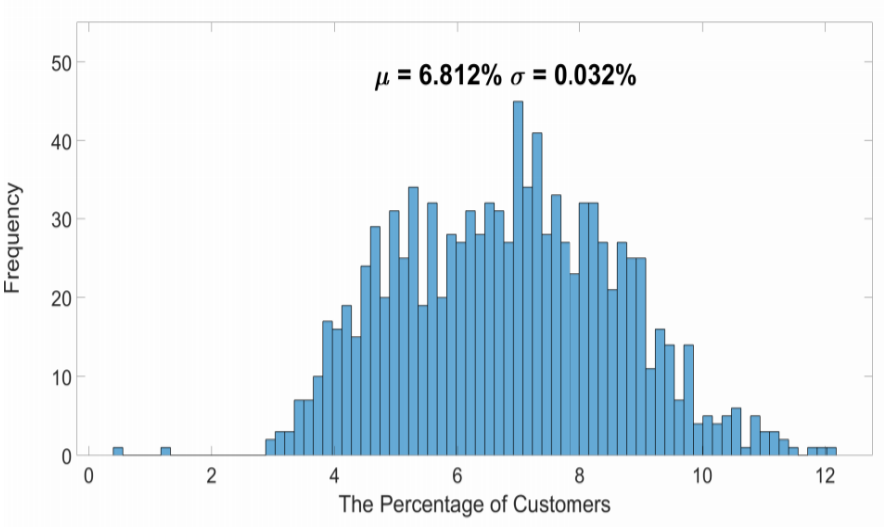
$$F_{j,m} = \frac{1}{n}\sum_{d=1}^{n}\frac{p_{j,m}(t_d)}{P_m(t_d)}$$

- $p_{j,m}(t_d)$- j-th customer's demand at system peak time on the d-th day of the m-th month
- $P_m(t_d)$- System peak demand on the d-th day of the m-th month
- t_d- System peak time on the d-th day of the m-th month

✓ **CMPC is basically the average customer contribution to the daily system peak demand during a month.**

Why not use customer peak load to represent customer peak contribution?

Evidence from real data:



The percentage of residential customers whose peak demand coincides with the system peak load (only around **6%** of customers have the same peak time as the system).

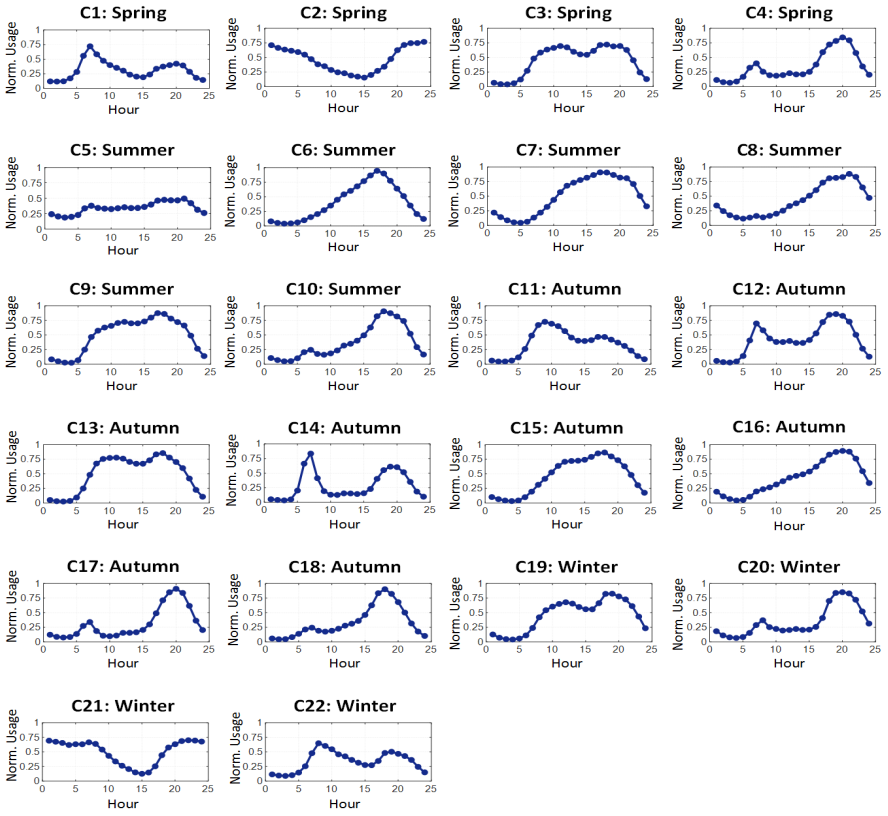
Step I: Seasonal Residential Customer Behavior Pattern Bank

- **Methodology**: unsupervised learning – spectral clustering (SC).
  - **Step I**: Obtaining seasonal average customer load profiles based on smart meter data.
  - **Step II**: Calculating pair-wise similarity to build a weight matrix, W, based on Gaussian kernel function.
  - **Step III**: Solving a graph partitioning problem. The objective function is to maximize both the dissimilarity between the different clusters and the total similarity within each cluster.
- **Two Main Advantages of SC**:
  - Utilizing the weight matrix of the dataset rather than using the high-dimensional demand profile data directly.
  - No assumptions on the data distribution.

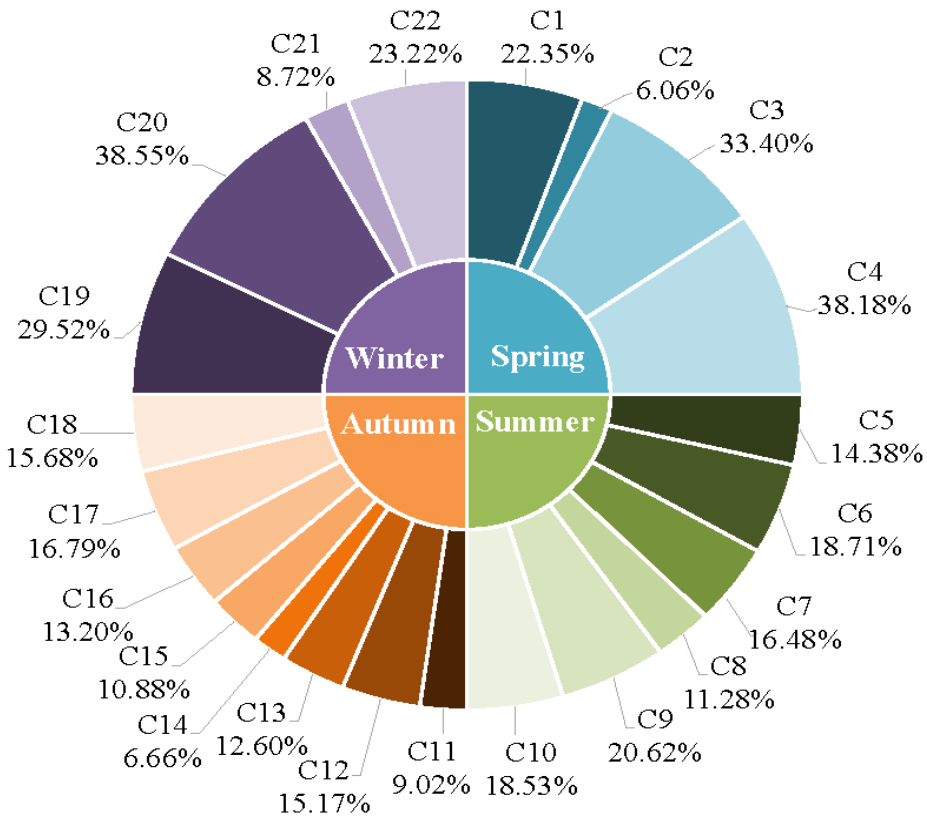
U. Luxburg, "A tutorial on spectral clustering," Statistics and Computing, vol. 17, no. 4, pp. 395–416, Mar. 2007.

Step I: Seasonal Residential Customer Behavior Pattern Bank

Typical discovered load profiles in different seasons from smart meter data



The percentage of observable customers belonging to each typical load profile in different seasons



Step II: Unobservable Customer Classification

- **Methodology**: supervised learning - multinomial logistic regression (MLR) algorithm (MLR).

- **Data**: The sociodemographic information of customers

- **Advantage of MLR**: MLR is able to obtain the likelihood of different typical profiles for customers rather than picking a single consumption pattern. The objective function of this classification model is defined as follows:

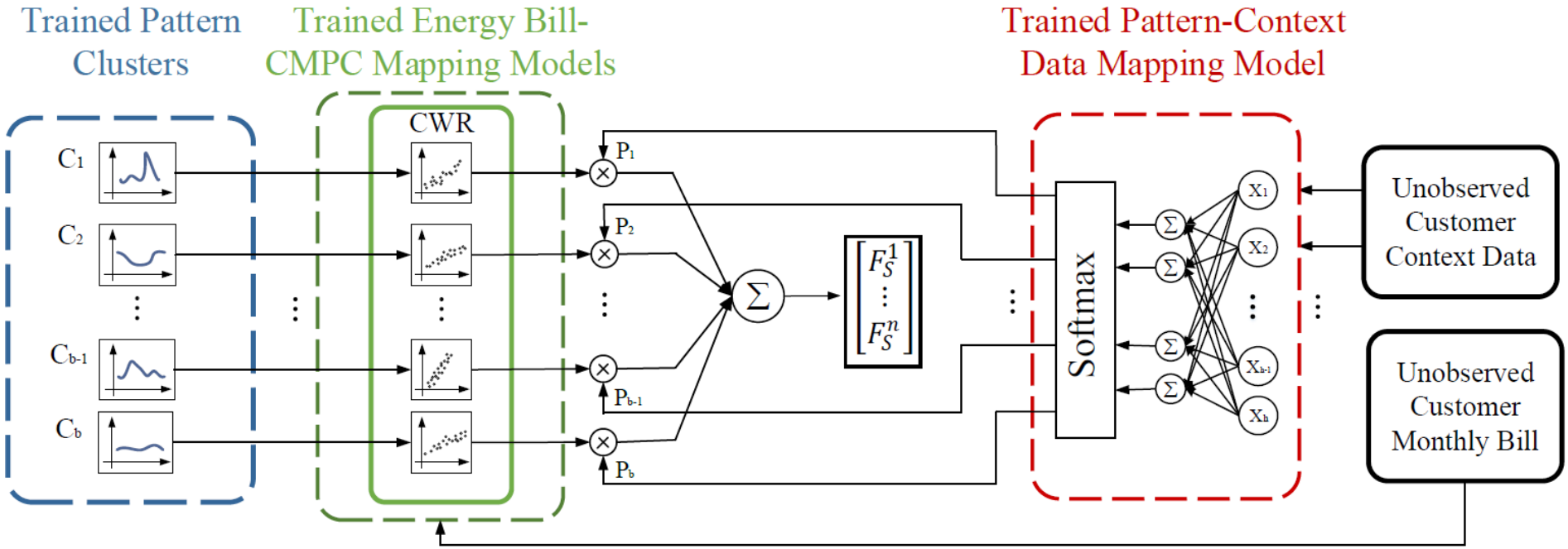
$$J = \sum_{j=1}^{M} [\sum_{z=1}^{k} c_j^z (w_z)^T X^j - log \sum_{z=1}^{k} \exp((w_z)^T X^j]$$

Where X^j is the approximate PDF of j'th customer, c_j^z is a binary string representing customer class membership.

Step III: Mapping CMPC with Energy Bills

- **Problem**: For observable customers in each pattern, what is the mapping relation between their CMPCs and monthly energy bills?
- **Data**: Monthly energy bills and CMPCs calculated using smart meter data.
- **Methodology**: supervised learning - linear regression model.
  - A linear regression model is assigned and trained for each typical load profile using the corresponding data.
  - The ordinary least square is utilized to estimate the parameters of the regression models.

Overview of Proposed Method



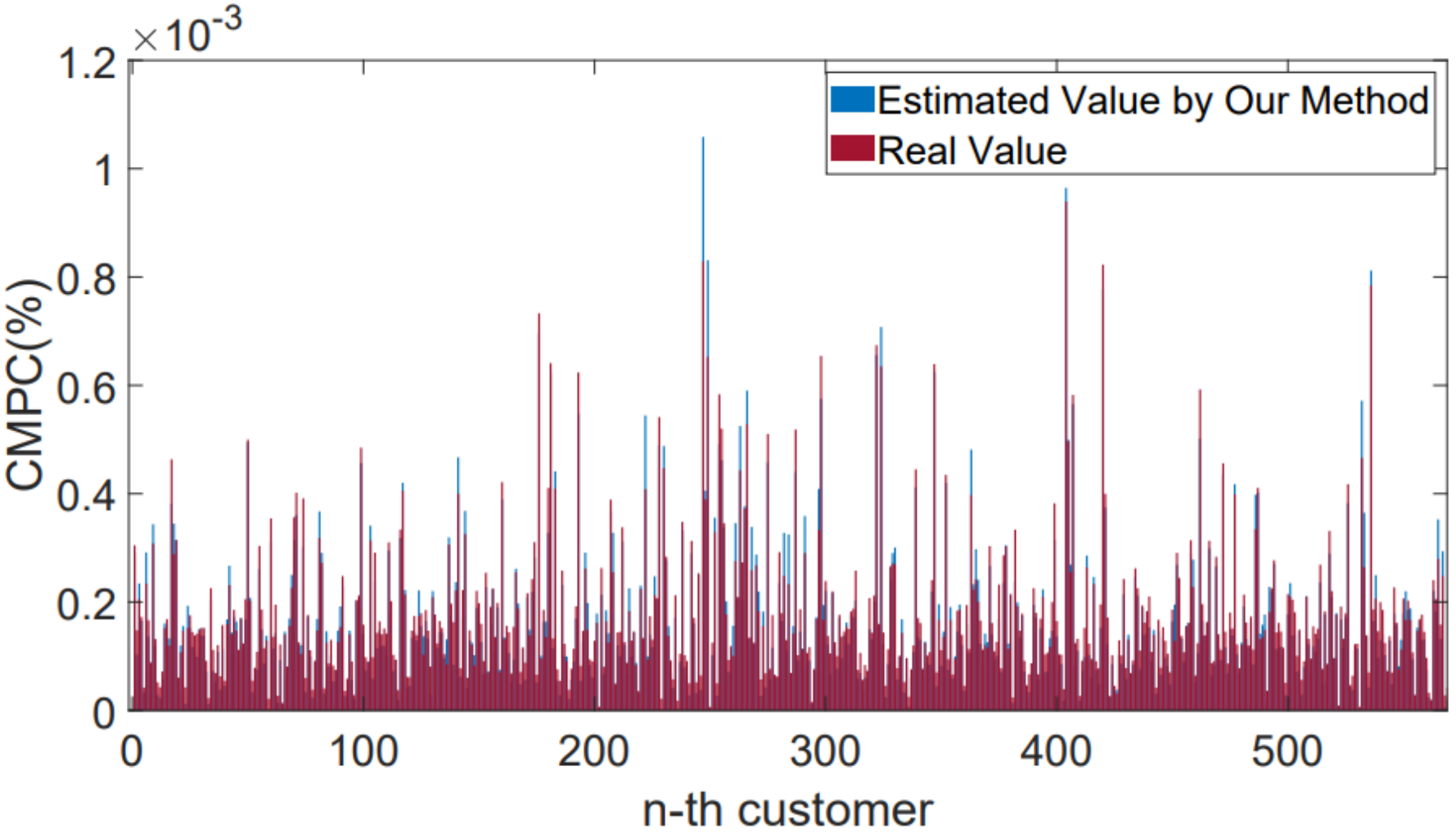
Ci – Typical Load Pattern
Pi – Classification Probability
Fi – Estimated CMPC

- ✓ **All regression models are merged into a weighted clusterwise regression model. The probabilistic outcomes of the classification model are assigned as the weight values.**

Numerical Results: Data Description

- The data includes the smart meter measurements of over **3,000** residential customers.
- The data ranges from **January 2015** to **May 2018**.
- The actual CMPC of each customer is calculated using real smart meter data.
- We assume that **20%** of customers are **unobservable** and then compare the estimated results with the actual CMPCs.

Numerical Results: Unobserved Individual Customer CMPC Inference

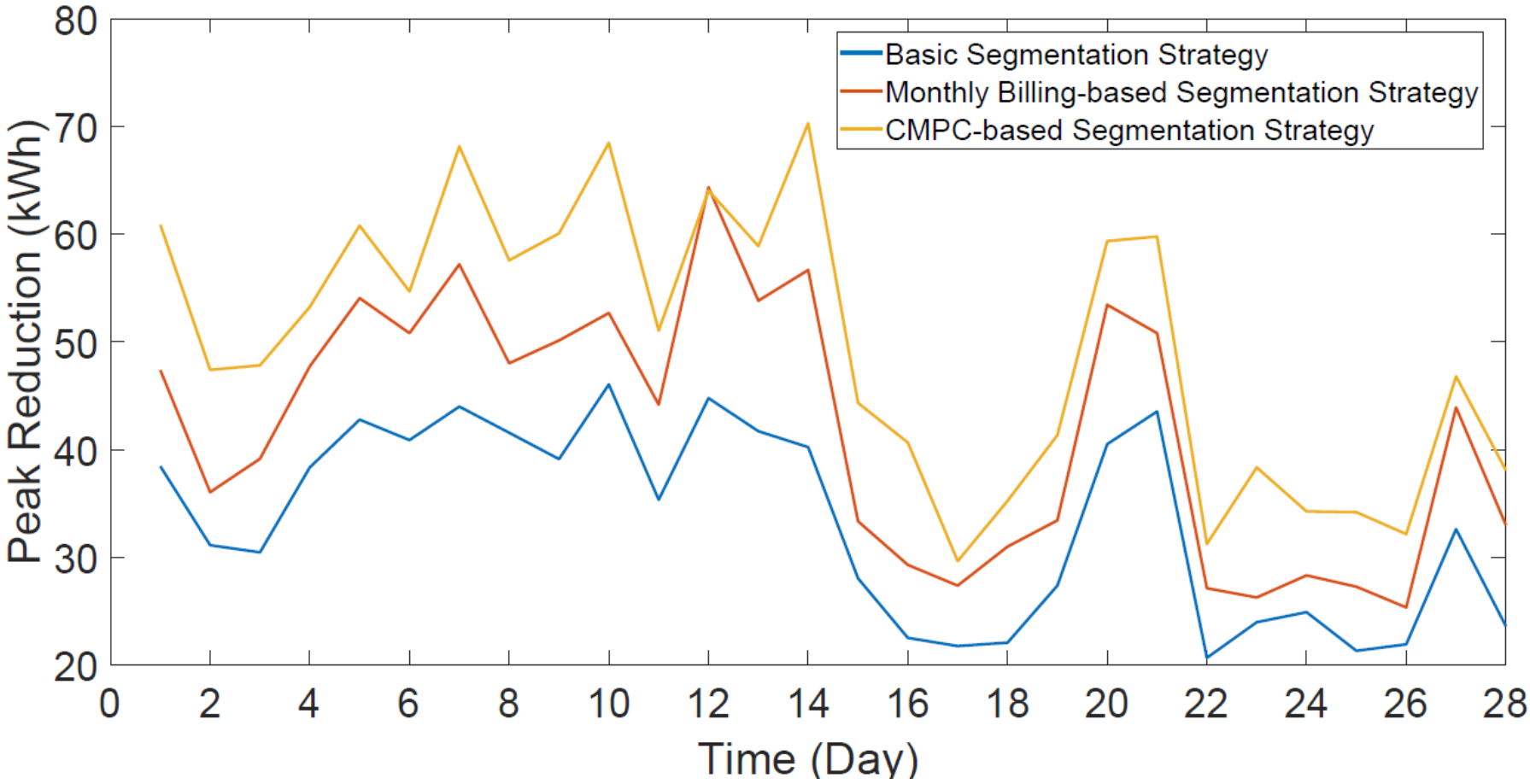


Numerical Results: Application of the Proposed Metric and Inference Method

- ✓ **Application:** select unobservable residential customers for the direct load control-based peak demand reduction.
- ✓ For a 300-house distribution system, 35% of residential customers are selected for meter installation and participation in peak shaving program.
- ✓ Three different customer segmentation strategies are tested: 1) select candidates randomly (base strategy); 2) select candidates by ranking monthly energy consumption; 3) select candidates based on the CMPC.
- ✓ We assumed the average load elasticity of customers to be 0.21 p.u.
- ✓ We have compared daily peak reductions for 28 days under the three different strategies.

C. Chen, J. Wang, and S. Kishore, "A distributed direct load control approach for large-scale residential demand response," IEEE Trans. Power Syst., vol. 29, no. 5, pp. 2219–2228, Sep. 2014.

Numerical Results: Application of the Proposed Metric and Inference Method



Smart Meter Data-Driven Outage Detection

- On August 10, a weather complex known as a “derecho” sent intense winds and thunderstorms over a 700-mile stretch in Midwest. Between August 10 and 13, total outaged customers were 1.9 million, with 585,000 in Iowa.
- The delay and inaccuracy of outage detection can cause waste of up to 80% of the invaluable restoration time.
- Conventional expert-experience-based methods that use customer calls are laborious, costly, and time-consuming.

Ames, Iowa, 8/11/2020

National Electrical Manufacturers Association, “Smart meters can reduce power outages and restoration time.” [Online]. Available: https: //www.nema.org/Storm-Disaster-Recovery/Smart-Grid Solutions/Pages/ Smart-Meters-Can-Reduce-Power-Outages-and-Restoration-Time.aspx

Outage Detection in Partially Observable Distribution Systems

Challenges:

- Smart meters can send last-gasp signals. However, most distribution systems are only **partially observable** (i.e., not every customer has smart meter).

- Most of the previous works handle the partially observable problem by involving extra data sources, such as real-time power-flow measurements and social network data.

- Outage detection can be considered as a **classification** problem (separating the data samples of normal and outage). However, the size of the outage data is far smaller compared to the data in normal conditions, which leads to a **data imbalanced problem**.

Outage Detection in Partially Observable Distribution Systems

Our Solution:

- ✓ Decomposing large-scale distribution networks into a set of intersecting **outage detection zones** and performing zone-based outage detection rather than branch-based outage detection.

- ✓ Optimizing the zone decomposition by exploiting the tree-like structure of distribution networks and the system observability (i.e., when system is fully observable, our method provides branch-based results).

- ✓ Developing an **unsupervised-based** model for outage detection (only utilize the data in normal conditions for model training).

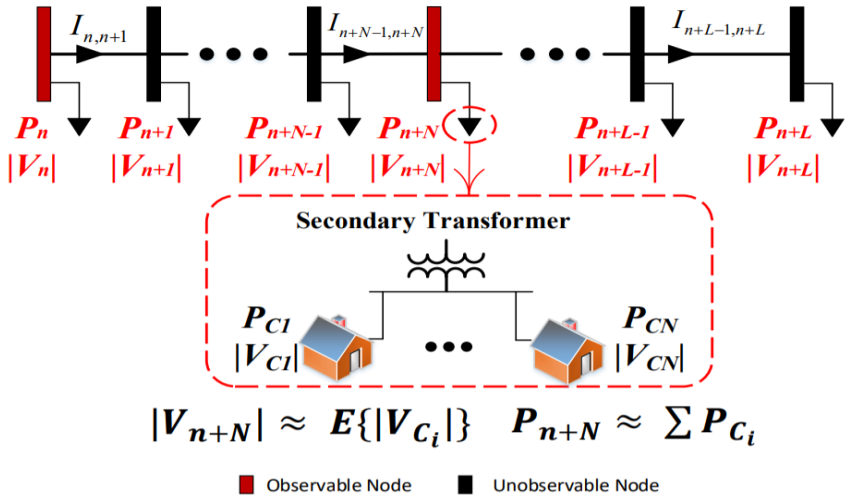
- ✓ Providing an anomaly score coordination process to accelerate outage location in large-scale networks.

Y. Yuan, K. Dehghanpour, F. Bu, and Z. Wang, "Outage detection in partially observable distribution systems using smart meters and generative adversarial networks, " IEEE Trans. on Smart Grid, accepted for publication.

Outage Detection Zone Definition

Definition: In a radial network, an outage detection zone, Ψ_i, is defined as $\Psi_i = \{S_{o1}, S_{o2}, Z_{\Psi_i}\}$, where S_{o1} and S_{o2} are two observable nodes, with S_{o1} being upstream of S_{o2}, and Z_{Ψ_i} is the set of all the branches downstream of S_{o1}.

- ✓ Give that an outage event anywhere in the zone will lead to deviations from the (voltage-power) data distribution obtained from two observable nodes under normal operations.



$$\Delta \mathbf{V_o} = |\mathbf{V}_n| - |\mathbf{V}_{n+N}| \approx \Delta \mathbf{V} + \sum_{i=n+1}^{\min(s,n+N)} \mathbf{K}_{i-1,i} \otimes \mathbf{I}_{i-1,i} \otimes \frac{\Delta \mathbf{P}_s}{cos\phi_s}$$

Step I: Breath-First Search-Based Zone Selection

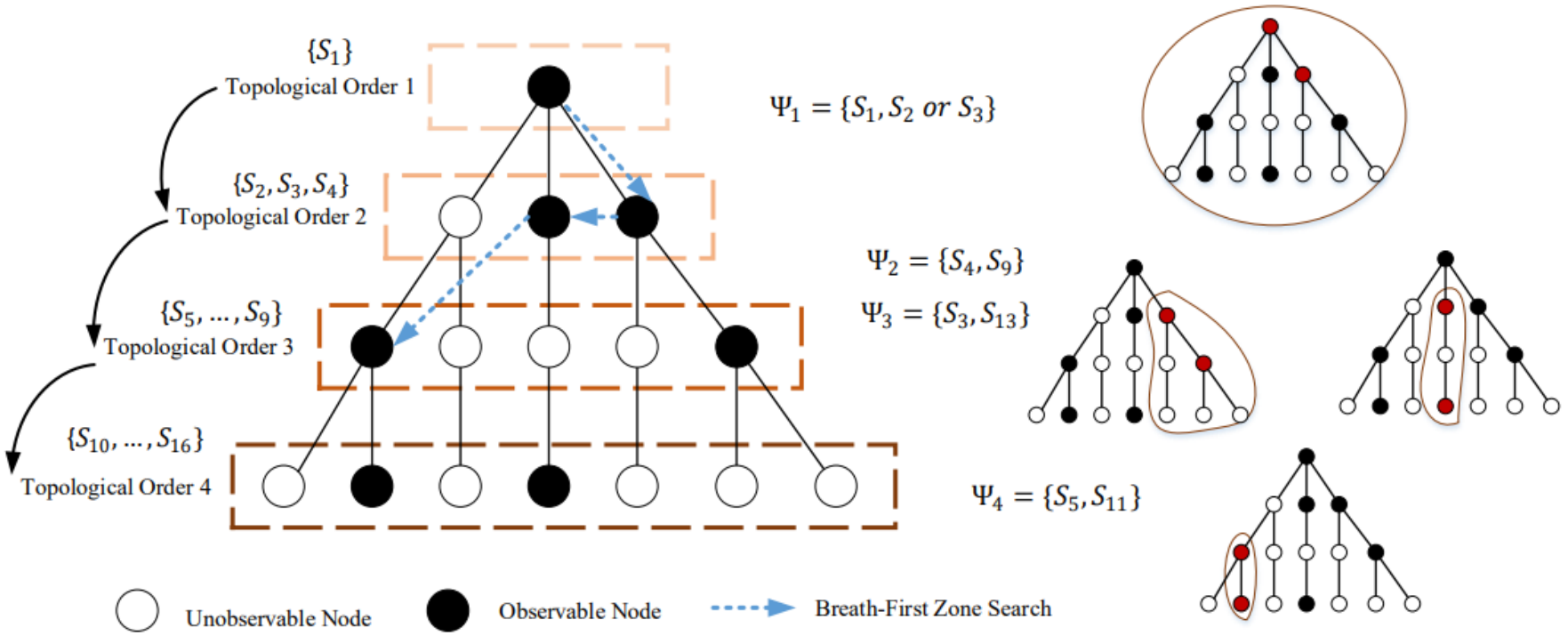
- **Problem**: How to optimally sectionalize networks into multiple zones based on the limited observability to maximize outage detectability?

- **Our Solution**: Proposing a breadth-first search-based mechanism to use **all observable node pairs** to build the zones.

  - Each branch in the system belongs to at least one zone.

  - Introducing a topological ordering, which simplifies outage location identification process.

Step I: Breath-First Search-Based Zone Selection



- Each zone is determined by two neighboring observable nodes and contains all branches downstream of these two nodes.
- Selecting the zones using observable nodes at the present layer before moving on the observable nodes at the next topological layer.
- The outcome of our zone selection algorithm follows a topological order, meaning that $\Psi_1 \succ \cdots \succ \Psi_w$.

Step II: Zone-Based Data Distribution Learning

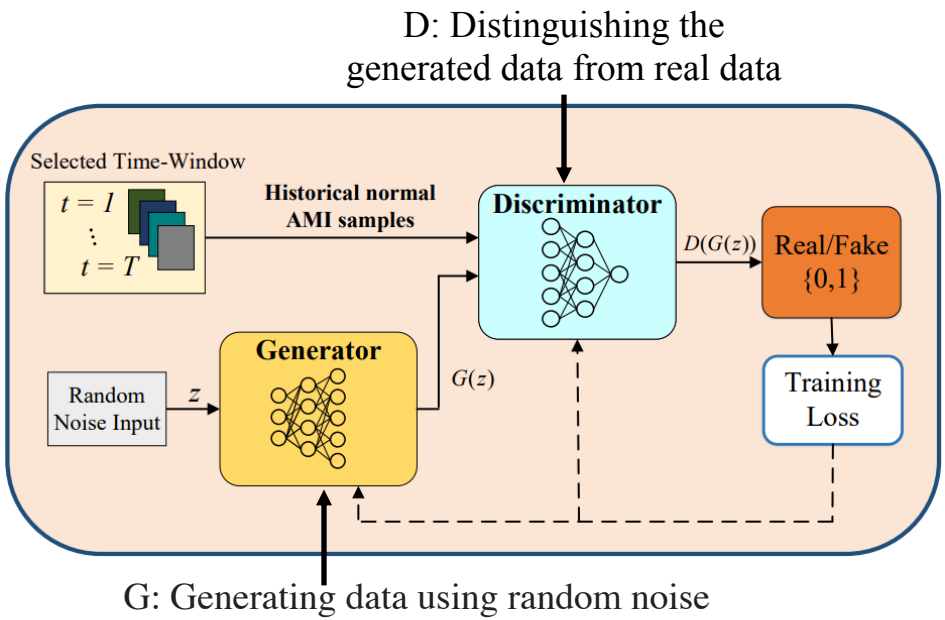
Challenge: Learning the distribution of measured variables $X = \{\Delta V^t, P_n^t, P_{n+N}^t\}_{t=1}^T$ within a time-window with length T (i.e., T = 3) for each zone (high-dimensional distribution).

Existing methods:

- Parametric-based methods require distributional assumptions.
- Traditional nonparametric-based methods (e.g., KDE) lack of scalability for large dataset.

Our Solution: Using Generative Adversarial Network (GAN) to implicitly and efficiently represent complex distributions without any distributional assumptions.

- To address data imbalanced problem, we only use the data in normal conditions.



Objective Function:

$$\min_{\theta_G}\max_{\theta_D} V(D,G) = \mathbb{E}_{x_{\Psi_i}\sim p_{x_{\Psi_i}}}(x_{\Psi_i})\left[\log\left(D(x_{\Psi_i})\right)\right] + \mathbb{E}_{z\sim p_z(z)}[\log\left(1-D(G(z))\right)]$$

Probability of D assigning the correct label to real samples.

Probability of D assigning the incorrect label to artificial samples from G.

Step III: Zone-Based Outage Detection

- Zone-based outage detection is achieved by defining a **GAN-based anomaly score** for each zone, which **quantifies deviations** between the learned normal data distribution and real-time measurements.

- The deviation is defined as follows:

$$\zeta_{\Psi_i}(x^t_{new}) = (1-\lambda)\cdot\delta_R(x^t_{new}) + \lambda\cdot\delta_D(x^t_{new})$$

δ_R is the **residual error** that describes the extent to which new measurement follows the learned distribution of the GAN:

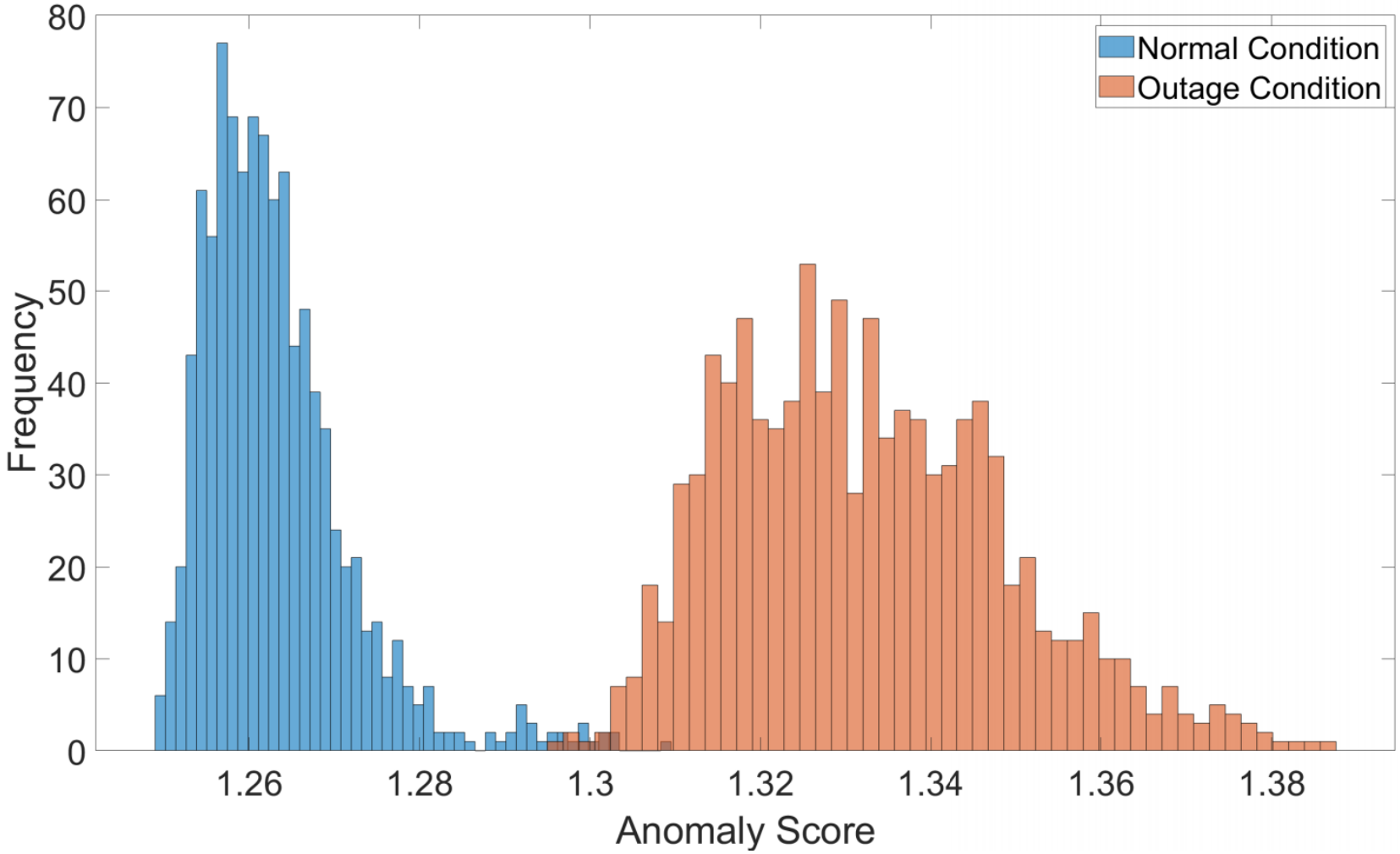
$$\delta_R(x^t_{new}) = \min_z |x^t_{new} - G(z)|$$

δ_D is the **discriminator error** that measures how well the optimal solution of the above optimization (z^*) follows the learned data distribution of the GAN.

$$\delta_D(x^t_{new}) = -\log D(x^t_{new}) - \log(1 - D(G(z^*)))$$

P. S. S. Thomas, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," IPMI, 2017.

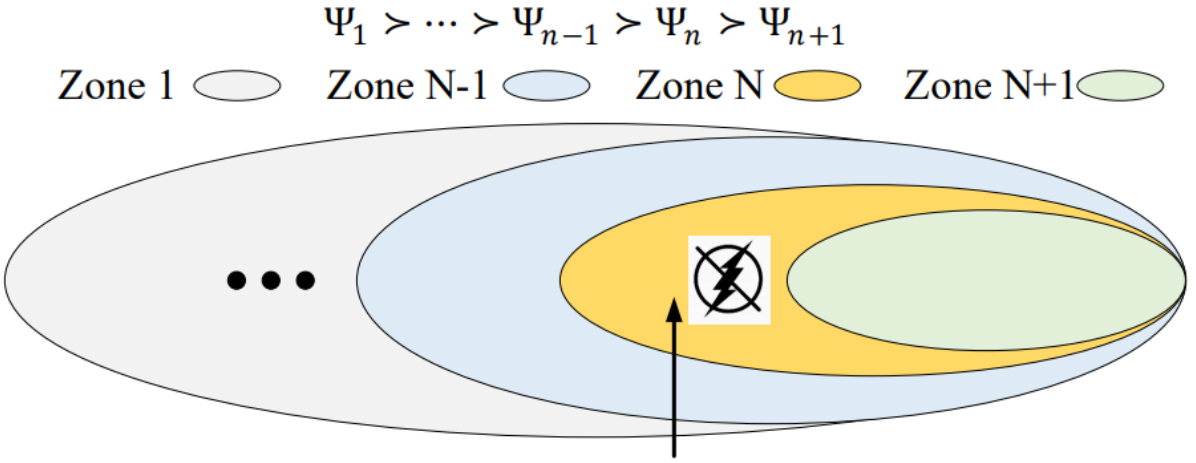
Step III: Zone-Based Outage Detection



✓ **A high anomaly score implies outage somewhere in the zone.**

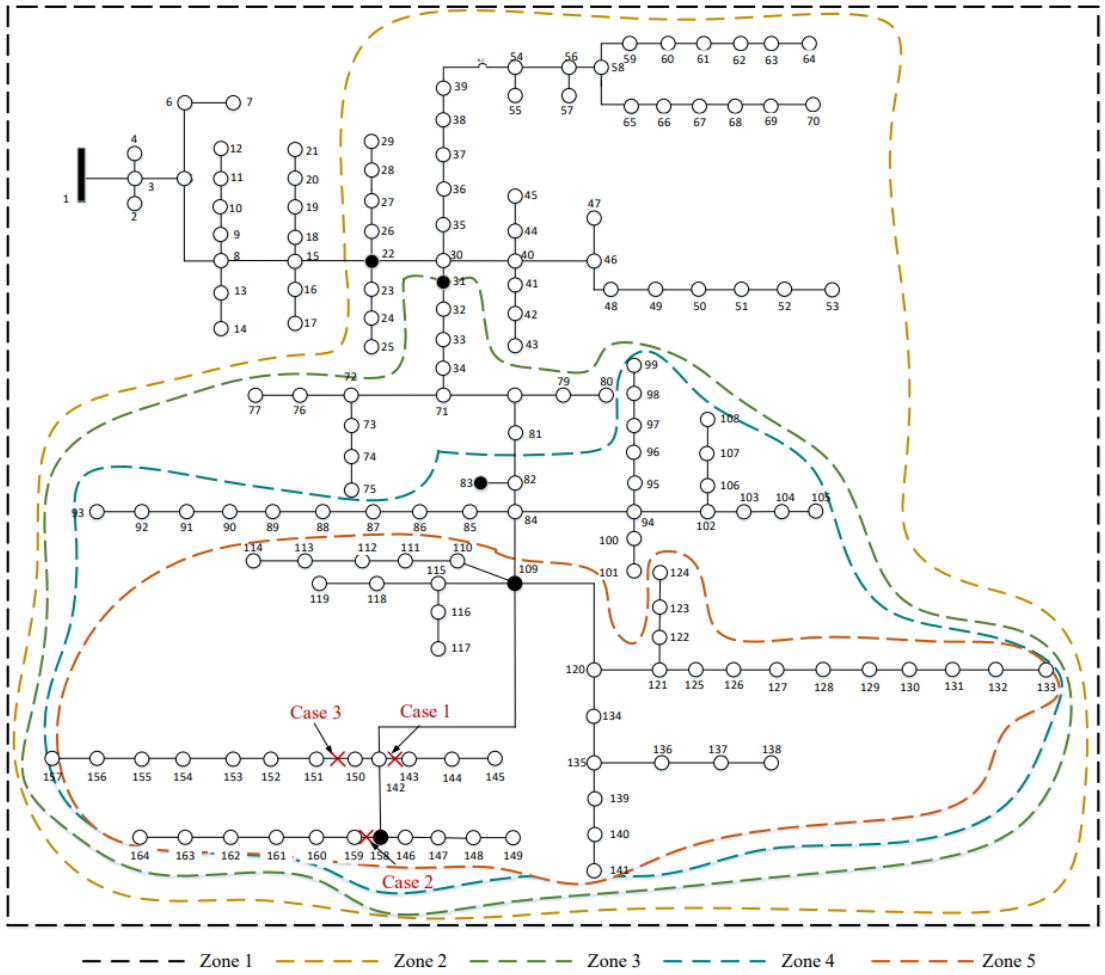
Step IV: GAN-Based Zone Coordination

- **Problem**: Multiple zones may contain the same outaged branch. How to down select the zone?
- **Solution**: Using the topological ordering and multiple anomaly scores.
- Zone coordination follows a bottom-up fashion until no outage-related zone exits.



Multiple zones include the outage location (i.e., Zone 1, Zone N-1, etc).
Zone N contains the maximum information on the outage event.
The minimum branch candidates are $Z_{\Psi_N} \backslash Z_{\Psi_{N+1}}$

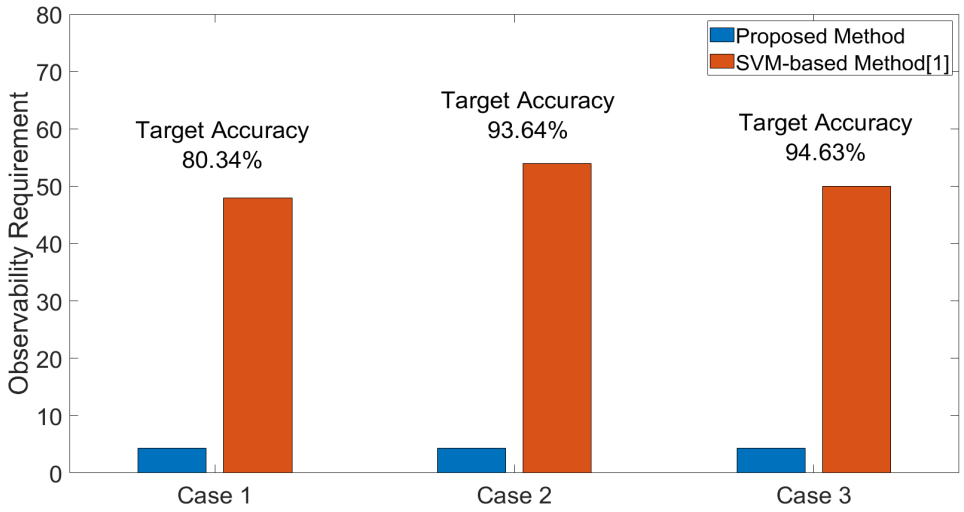
Numerical Results: 164-node Feeder Topology



- Six observable nodes are assumed in this feeder (Node 1, 22, 31, 83, 109, 158).
- Five zones are defined based on these nodes $\Psi_1 \succ \Psi_2 \succ \Psi_3 \succ \Psi_4 \succ \Psi_5$.
- Three outage events are simulated with different outage magnitudes (**case 1**: 20 customers are disconnected; **case 2**: 50 customers are disconnected; **case 3**: 80 customers are disconnected.)

Numerical Results: Accuracy Analysis

	Outage Detection Accuracy
Case 1	80.34%
Case 2	93.64%
Case 3	94.63%

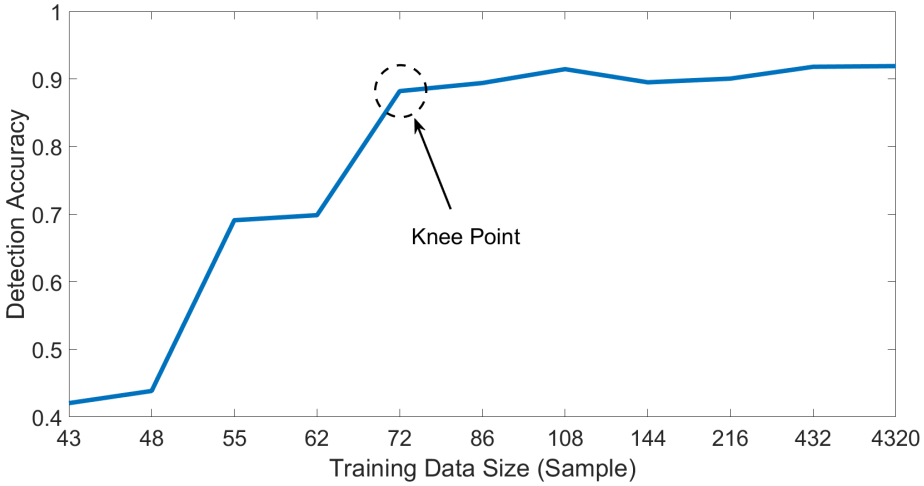


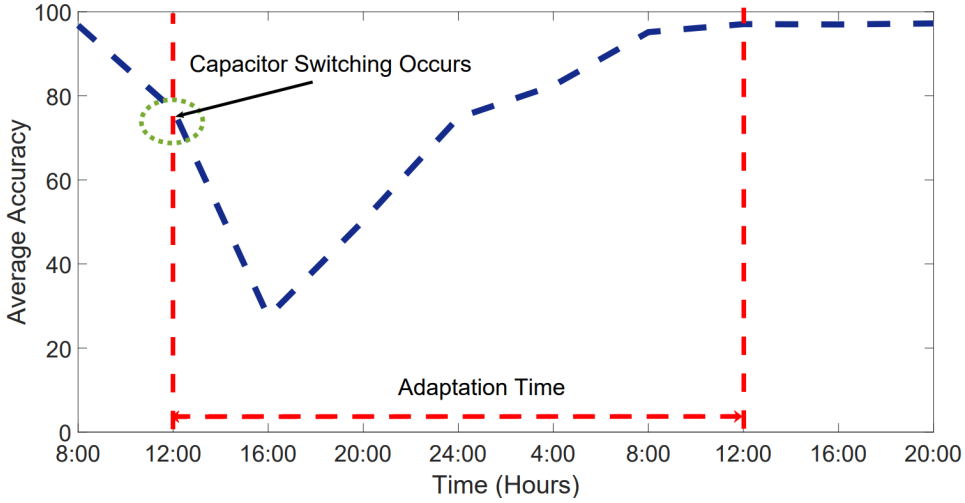
- ✓ **For three cases, we have tested if our method can detect outages in zone 5. The table shows the results for three cases.**
- ✓ **We have conducted numerical comparisons with a previous method.**

- ✓ **A previous method uses the last gasp signal from smart meters as the input of SVM to identify event location.**
- ✓ **The previous method requires a much higher level of observability (i.e., around 10 times) to achieve similar accuracy as our method.**

Z. S. Hosseini, M. Mahoor, and A. Khodaei, "AMI-enabled distribution network line outage identification via multi-label SVM," IEEE Trans. Smart Grid, vol. 9, no. 5, pp. 5470–5472, Sep. 2018.

Numerical Results: Sensitivity Analysis and Method Adaption





- ✓ **The performance of our model can reach reasonable detection accuracy with a small training set (around 3 days of data, hourly smart meter data).**
- ✓ **Our method can adapt to changes in system conditions (i.e., capacitor switching) with a relatively short time (around 1 day).**

Conclusion and Future Work

- Smart meter data, although may be of low resolution and limited measurement variables, can be used to significantly enhance distribution system observability. There are many applications such as load profiling, outage detection, behind-the-meter solar disaggregation and network modeling.

- However, many utilities do not have full smart meter coverage. We demonstrated how to use available smart meter data together with machine learning to estimate unobservable customers' peak contributions and detect outages in partially observable systems.

- In the future, we will focus on using real smart meter data to identify/calibrate network models as well as distribution system state estimation.

Distribution Course Material Sharing

EE653: Power distribution system modeling, optimization and simulation

- Introduction to Distribution Systems
- Modeling Series Components – Distribution Lines
- Modeling Series Impedance of Overhead and Underground Lines
- Modeling Shunt Admittance of Overhead and Underground Lines
- Modeling Shunt Components – Loads and Caps
- Distribution Feeder Modeling and Analysis Part I
- Modeling Voltage Regulators
- Modeling Three-Phase Transformers
- Distribution Feeder Modeling and Analysis Part II
- Various Power Flow Calculation Methods in Distribution Systems
- Optimal Power Flow in Distribution Systems
- Voltage/VAR Optimization and Conservation Voltage Reduction
- Distribution System State Estimation and Smart Meter Data Analytics
- Microgrids – Introduction and Energy Management
- Microgrids – Dynamic Modeling and Control
- OpenDSS Tutorial
- Real Distribution System Modeling and Analysis using OpenDSS
- Introduction to GridLAB-D
- Distribution System Resilience: Hardening, Preparation and Restoration
- Energy Storage

- You may download the course material at: http://wzy.ece.iastate.edu
- All slides are editable, feel free to use.
- Comments are very welcome!

Thank You!
Q & A

Zhaoyu Wang
http://wzy.ece.iastate.edu