# Voltage Sag Prediction Model Using Distribution Power Quality Phase II Data

*Technical Report*

# Voltage Sag Prediction Model Using Distribution Power Quality Phase II Data

**1002199**

Final Report, December 2003

EPRI Project Manager
A. Sundaram

EPRI • 3412 Hillview Avenue, Palo Alto, California 94304 • PO Box 10412, Palo Alto, California 94303 • USA
800.313.3774 • 650.855.2121 • askepri@epri.com • www.epri.com

## DISCLAIMER OF WARRANTIES AND LIMITATION OF LIABILITIES



ORGANIZATION THAT PREPARED THIS DOCUMENT

**EPRI PEAC Corporation**

## ORDERING INFORMATION

Requests for copies of this report should be directed to EPRI Orders and Conferences, 1355 Willow Way, Suite 278, Concord, CA 94520, (800) 313-3774, press 2 or internally x5379, (925) 609-9169, (925) 609-1310 (fax).

# CITATIONS

This report was prepared by

EPRI PEAC Corporation
942 Corridor Park Blvd.
Knoxville, TN 37932

Principal Investigators
T. Short
W. Sunderman
A. Maitra

This report describes research sponsored by EPRI.

The report is a corporate document that should be cited in the literature in the following manner:

*Voltage Sag Prediction Model Using Distribution Power Quality Phase II Data*, EPRI, Palo Alto, CA: 2003. 1002199.

# PRODUCT DESCRIPTION

The large amount of data gathered in EPRI's Distribution Power Quality (DPQ) Phase II study provides an opportunity to explore ways to predict voltage sags at locations without monitoring. Each of the DPQ Phase II sites has a number of site characteristics that can be used in a statistical model to predict voltage sags. Such predictions can help utilities in many ways. Predictions and prediction limits can define risks associated with performance-based contracts. Utilities can use predictions to decide whether a given circuit area is under-performing relative to other sites with similar characteristics. Utilities can pass on expected performance numbers to their customers to help them fortify their facilities.

## Results & Findings

Researchers combined the DPQ Phase II database with the database from the original DPQ study to form a combined database from which prediction models for voltage sags could be developed. The researchers analyzed the variability and uncertainty of various models and derived a relatively straightforward set of equations from the data for the major SARFI (System Average RMS [variation] Frequency Index) voltage-sag indices.

## Challenges & Objectives

The main objective of this study was to develop and test prediction models for voltage sags based on the DPQ Phase II dataset. These are statistical regression models built from optimizations of models containing several of the site characteristics available for the DPQ sites. The main challenges were finding good models, dealing with sites that had unknown site characteristics, and analyzing the variability in the models.

## Applications, Values & Use

This report provides utility engineers with a set of straightforward equations for estimating the rate of voltage sags at a location. Only a few site parameters are needed: substation transformer size, number of feeders on the bus, distribution voltage level, total feeder length, lightning ground flash density, and a crude estimate of tree coverage. The prediction model also includes an estimate of variability, which can be portrayed as prediction limits.

## EPRI Perspective

Estimating the frequency and characteristics of power quality events, which are part of the normal electrical environment, is critical for designing immunity into end-use equipment. This information is also important to define baseline levels of power quality that can be expected by a customer and sets the stage for providing premium power services. The prediction models developed with the DPQ II data will help utilities realize the goal of enhancing compatibility of sensitive equipment with the electrical environment.

## Approach

The main components of the prediction modeling are model fitting, variable selection, handling missing data, and quantifying variability. The best prediction models were linear with a Gamma error distribution. The variables most impacting SARFI were tried in various model formulations. The DPQ I/DPQ II dataset has a number of sites where some of the site characteristics are unknown. The researchers used a technique known as “multiple imputation” to fill in missing data in a manner that allowed them to make predictions without the missing data distorting the predictions. Variability is an important result of a prediction. Bootstrapping was explored as a way of improving prediction limits.

## Keywords

Power quality
Distribution
Voltage sag
Power quality monitoring

# ACKNOWLEDGMENTS

The authors would like to acknowledge the help of Tim Hesterberg of the Insightful Corporation, the makers of the S-PLUS statistical analysis software. Tim provided significant insight into bootstrapping statistical regression models and using a Bayesian approach to incorporate measurement data into prediction models.

The authors thank the following individuals for their contributions and assistance throughout the DPQ phase II project, in providing both measurement data and site characteristic data, which were critical to this project:

| | |
|---|---|
| Austin Energy | Mehrdad Vatani |
| | Jim Blackwell |
| | Robert Burns |
| BGE | Vladimir Basch |
| CLP Power | Vincent Kwok Wing To |
| Carolina Power & Light (a Progress Energy Company) | Stephen Middlekauff |
| | Glenn Lampley |
| Dominion Virginia Power | Bobby Gettier |
| | Noel Freeman |
| Duke Power | Patrick Chaney |
| | Steve Whisenant |
| Pacific Gas and Electric | Gilbert Hensley |
| Public Service Electric and Gas | Greg Olson |
| San Diego Gas and Electric | Henry “Rusty” Nelson |
| | Robert J. Gilleskie |
| | Joe Frani |

| | |
|---|---|
| Salt River Project | Rao Thallam<br>Kevin Kittredge<br>Steve Sturgill<br>John Blevins |
| Tennessee Valley Authority | Fred Elmendorf<br>Theo Laughner<br>Michael Ingram |
| United Illuminating | Marek Waclawiak |

This report contains material reprinted with permission from:

T. A. Short, *Electric Power Distribution Handbook*, CRC Press, Boca Raton, FL, 2004.

# EXECUTIVE SUMMARY

## Project Overview

EPRI's Distribution Power Quality (DPQ) Phase II effort in 2002 resulted in a database of voltage-sag information at numerous customer sites, including transmission, distribution, and secondary networks. In 2003, data were obtained from DPQ Phase II along with data from the original DPQ study (now referred to as DPQ Phase I) to develop a linear model that predicts the voltage-sag rate at a particular site, given the site characteristics. Predictions are provided for all of the major SARFI (System Average RMS [Variation] Frequency Index) indices.

## Variables Most Correlated With Voltage Sags

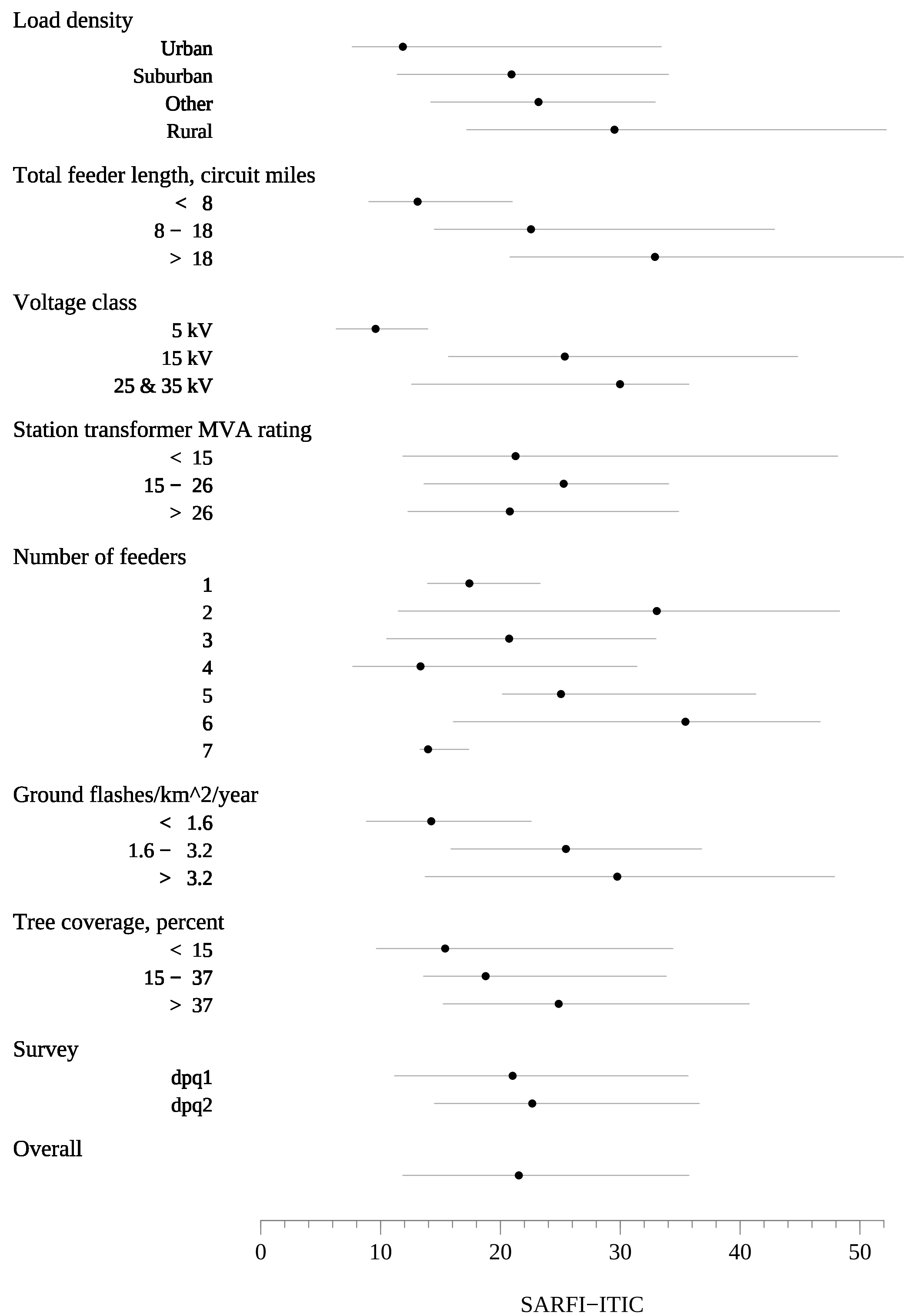
Several site characteristics are available for the DPQ I and DPQ II monitoring locations. Many of the relationships between site characteristics on voltage sags as characterized by $SARFI_{ITIC}$ are shown in Figure ES-1. Many of these trends are expected. Urban sites tend to have fewer sags than suburban or rural sites. Systems that operate at higher nominal voltages have more sags than lower-voltage systems. Circuits in areas with higher lightning have more voltage sags. No one site characteristic dominates. The most pronounced relationships are load density, system voltage, and lightning activity.

Keep in mind that the sites have high variability and significant overlap within each category. Figure ES-1 has bars that range between the lower and upper quartiles, so half of the site measurements within the given category fall within the bands shown.

## Prediction Procedures

The main components of the prediction modeling are:

- *Model fitting* – The best prediction models were linear with a Gamma error distribution. The variables most impacting SARFI were tried in various model formulations.
- *Missing data* – The DPQ I/DPQ II dataset has a number of sites where some of the site characteristics are unknown. We used a technique known as “multiple imputation” to fill in missing data in a manner that allowed us to make predictions without the missing data distorting the predictions.
- *Variability* – Variability is an important result of a prediction. We explored bootstrapping as a way of improving prediction limits. Fortunately, bootstrapping is not generally needed.




The dots mark the site median, and the bands show the range between the upper and lower quartiles.

**Figure ES-1**
**$SARFI_{ITIC}$ at Feeder Locations Grouped by Several Different Characteristics**

## Predictions

The main result of this analysis is the SARFI predictions, which fit the following equation with the coefficients given in Table ES-1:

$$SARFI_X = k_0 + k_1 l + k_2 N_g + k_3 \frac{n_f \cdot kV^2}{MVA_{xfmr}}$$
$$+ k_4 \text{ if moderate to heavy tree coverage}$$

where,

> $SARFI_x$ = predicted annual number of events that sag below the given SARFI criteria
>
> $l$ = total exposure (including three-phase and single-phase portions) on the circuit in miles
>
> $N_g$ = lightning ground flash density in flashes/km$^2$/year
>
> $kV$ = base line-to-line voltage in kV
>
> $n_f$ = total number of feeders off the substation bus
>
> $MVA_{xfmr}$ = station transformer base rating (open-air rating) in MVA

**Table ES-1**
**Coefficients for Prediction Equations for Various SARFI Indices**

| | $k_0$ Intercept | $k_1$ $l$ | $k_2$ $N_g$ | $k_3$ $\frac{n_f \cdot kV^2}{MVA_{xfmr}}$ | $k_4$ Tree term | Dispersion |
|---|---|---|---|---|---|---|
| $SARFI_{80}$ | 11.99 | 0.426 | 2.71 | 0.322 | 3.55 | 0.399 |
| $SARFI_{ITIC}$ | 4.26 | 0.480 | 2.20 | 0.247 | 6.33 | 0.389 |
| $SARFI_{70}$ | 6.23 | 0.353 | 2.24 | 0.195 | 4.37 | 0.435 |
| $SARFI_{SEMI}$ | 3.50 | 0.356 | 1.56 | 0.135 | 6.18 | 0.436 |
| $SARFI_{50}$ | 4.33 | 0.358 | 1.64 | 0.000 | 3.89 | 0.552 |

Keep in mind that predictions from these formulas are not precise estimates, but they do help more accurately define the probability range at a given site. The dispersion term of the model can be used to find prediction limits from the Gamma distribution. The variability in the model prediction includes three main components: time variability, site-to-site variability, and model uncertainty. It is difficult to separate these effects from each other.

To judge the precision and effectiveness of these models, refer to Figure ES-2 for predictions of $SARFI_{ITIC}$. While there is considerable variability, the model does well in quantifying the variability over the range of values in the dataset.

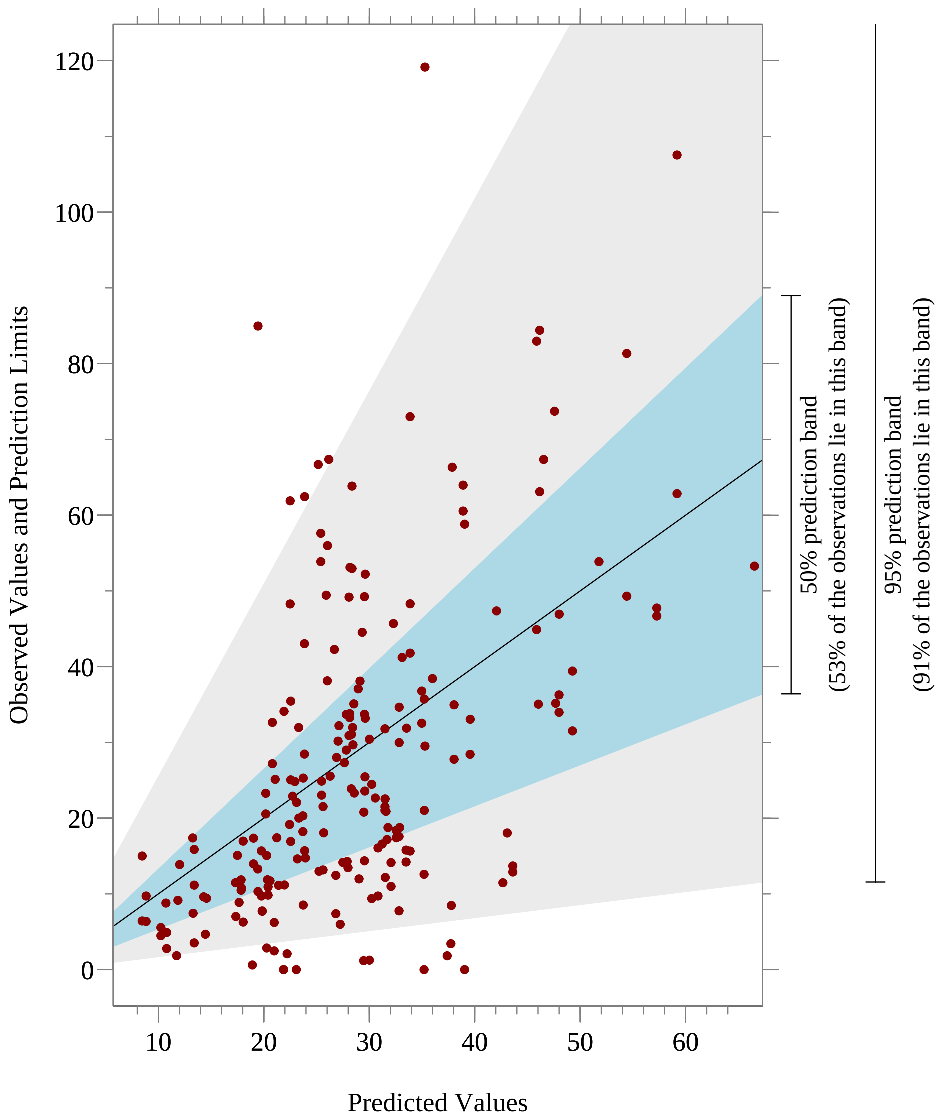



**Figure ES-2**
**Observed $SARFI_{ITIC}$ Versus Predicted $SARFI_{ITIC}$ for DPQ I/DPQ II Feeder Site Data**

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# 1
# PROJECT OVERVIEW

## Introduction

EPRI's Distribution Power Quality (DPQ) Phase II effort in 2002 resulted in a database of voltage-sag information at numerous sites, including transmission, distribution, and secondary networks. This year's project uses data obtained from DPQ Phase II to develop a linear model that predicts the voltage-sag rate at a particular site, given the site characteristics.

Voltage-sag predictions can help utilities in many ways. Predictions and prediction limits can define risks associated with performance-based contracts. Utilities can use predictions to decide if a given circuit area is under-performing relative to other sites with similar characteristics. Utilities can pass on expected performance numbers to their customers to help them fortify their facilities.

## Project Objective

The project used data from DPQ Phase II to identify the correlation of site characteristics to voltage sag rates, using statistical techniques to evaluate correlation coefficients for each site variable. From this analysis, it we chose a set of site descriptors and developed a linear sag predictor model using a multiple regression analysis technique.

## Background on the DPQ Studies

In the early 1990s, EPRI initiated a project called Distribution Power Quality (DPQ), which resulted in power quality monitoring at 277 distribution sites, which were statistically selected throughout the United States, to gain valuable knowledge regarding the frequency and severity of power quality events. The data collected for a period of more than two years were analyzed to understand the frequency and severity of different types of power quality events. The data from DPQ—called DPQ I in this report—were used to compose the first and only comprehensive database that characterizes the power quality level in distribution systems.

Since the completion of DPQ I in 1995, several utilities have implemented system-wide power quality monitoring programs in both distribution and transmission systems. The wealth of data collected by these utilities provides a unique opportunity to synthesize meaningful information regarding the variability of grid power based on system characteristics, especially information regarding the rate of voltage sags. It also presents a unique opportunity to compare the results of DPQ I with the data from these monitoring programs.

In 2001, there were two primary focuses for EPRI's Premium Power Grades Project. The first was to identify the technical characteristics of distribution and transmission sites with permanently installed power quality monitors. From that survey, two characteristics were selected to stratify the monitor sites: voltage class and lightning flash density. The second focus of the Premium Power Grades Project was to develop site-selection criteria for sampling the available utility monitoring data. This sample would eventually be used to represent the population of monitored sites.

Based on sampling data, the volunteer utilities randomly selected sites in each voltage class and lightning-flash-density strata. From this group of sites, the utilities provided more detailed site characteristics and the actual power quality data. The power quality data were delivered in several different formats. The data were then analyzed in a manner similar to DPQ I.

For more detailed information on the DPQ I study, refer to:

- *An Assessment of Distribution System Power Quality: Volume 2: Statistical Summary Report*, EPRI, Palo Alto, CA: 1996. 106294-V2.
- *An Assessment of Distribution System Power Quality: Volume 3: Library of Distribution System Power Quality Monitoring Case Studies*, EPRI, Palo Alto, CA: 1996. 106294-V3.

For more information on the DPQ II study, refer to:

- *Premium Power Grades: Developing Site Selection Criteria for DPQ Phase II*, EPRI, Palo Alto, CA: 2002. 1005921.
- *Distribution System Power Quality Assessment: Phase II: Voltage Sag and Interruption Analysis*, EPRI, Palo Alto, CA: 2003. 1001678.

## Important Distinctions Between DPQ I and DPQ II

There are several important points on which DPQ I and DPQ II differ. The most important distinction between the two studies is the population from which the monitored sites were chosen. DPQ I selected sites from a population of distribution feeders and substations where no monitors were installed. The monitors that collected data for DPQ I were installed after site selection and during the project at only distribution locations. On the other hand, DPQ II selected sites where monitors were already installed by participating utilities at both transmission and distribution locations. Additionally, the following points should be taken into consideration in the interpretation of the results presented in this report:

- Monitor types
- Monitor settings
- System levels
- Scope

### *Monitor Types*

There was only one type of monitor used during DPQ I. However, 11 different types monitors were used during DPQ II. Different types of monitors record event data differently depending on sampling rates, triggering algorithms, and data storage. For example, a relay that samples a waveform at four samples per cycle versus a dedicated power quality monitor that samples a waveform at 128 samples per cycle may report slightly different event durations.

### *Monitor Settings*

During DPQ I, the monitor settings for measuring root-mean-square (rms) voltage variation were tightly controlled. However, the triggering thresholds and recording parameters for the monitors used in DPQ II were set prior to the initiation of the project. The participating utilities set the triggering thresholds of their monitors based on several, sometimes conflicting, objectives, such as adhering to international standards, adhering to state reporting requirements, meeting operational requirements, or trying to closely match key customer sensitivity.

### *System Locations/Levels*

During DPQ I, selected sites fell into three system-level location categories: locations SM, MF, and EF shown in Figure 1-1. Sites in all three location categories were within the distribution system. However, during DPQ II, selected sites fell into eight location categories: locations A through G shown in Figure 1-2 plus another category (H), which indicates a location unspecified by the participating utility. Also, sites in location categories C and D were within the transmission system. Therefore, DPQ II includes data from both the distribution and transmission system. Because the population of sites in DPQ II already had installed monitors, it is likely that these sites serve key customers and critical transmission or distribution lines.

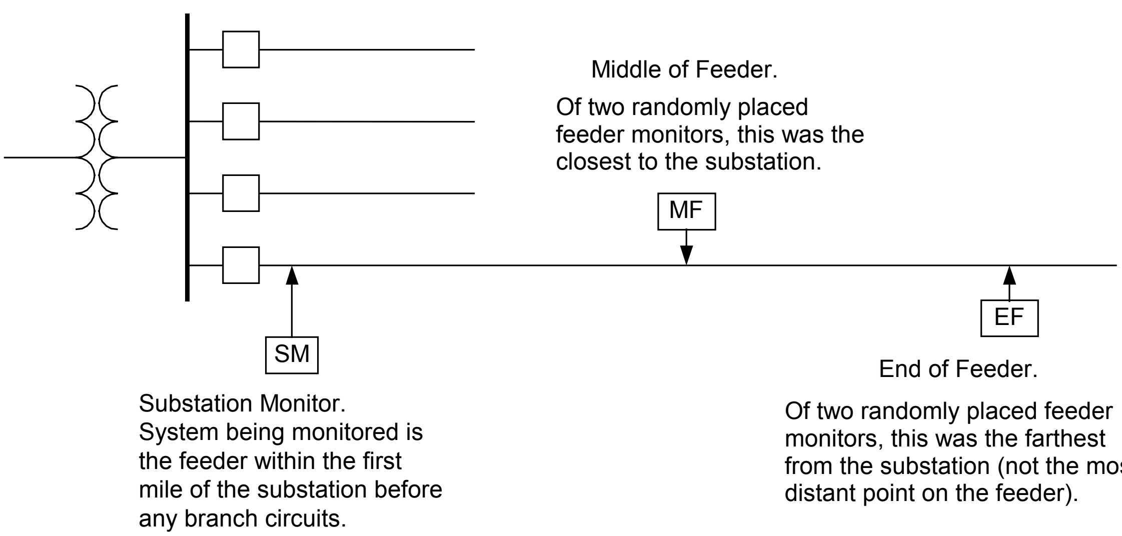



**Figure 1-1**
**Location Categories for the Sample Sites of DPQ I**

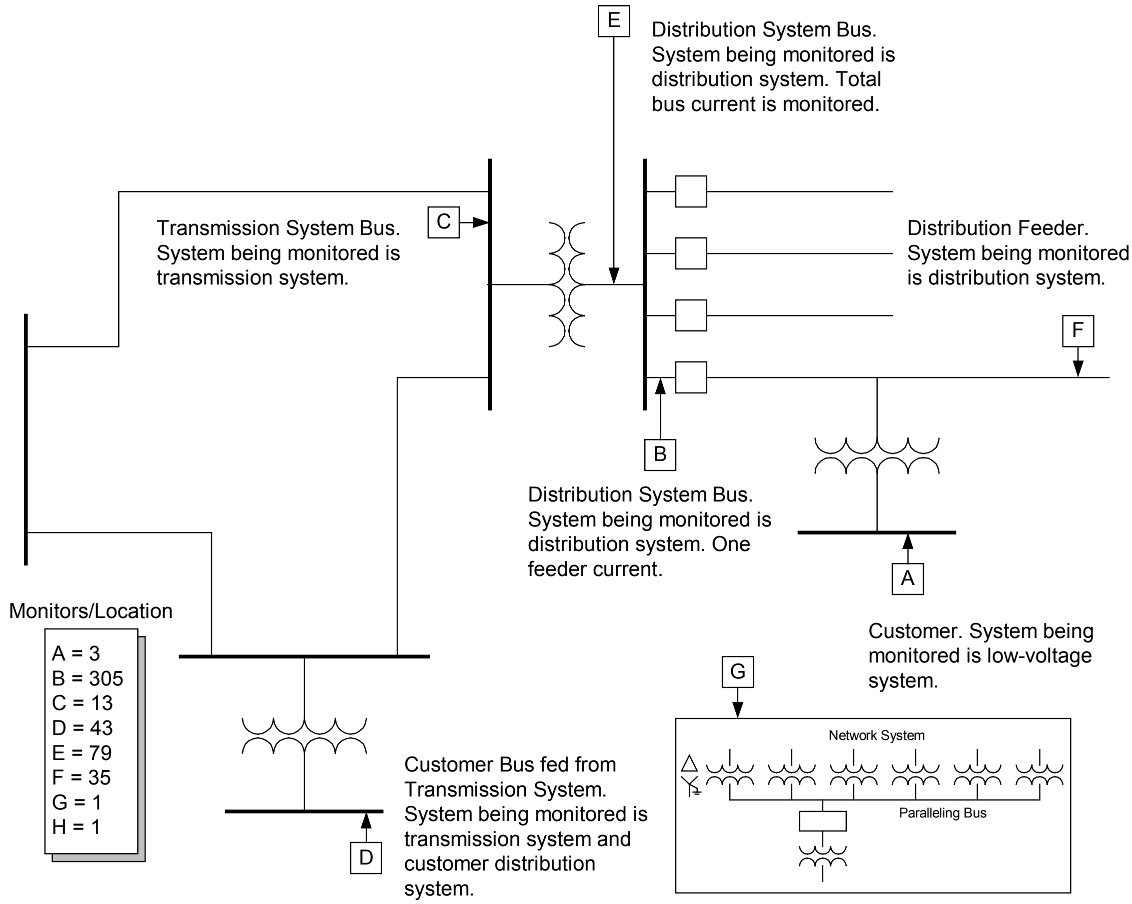



**Figure 1-2**
**Location Categories for the Sample Sites of DPQ II**

## *Scope*

The data-collection portion of DPQ I, which covered RMS variation, was conducted over a two-year time frame and included 277 sites. However, the data-collection portion of DPQ II spanned nearly a decade and included 480 sites. Figure 1-3 shows the available date range for each of the 13 participating utilities.

Table 1-1 highlights some of the key differences in the overall scope of each study.

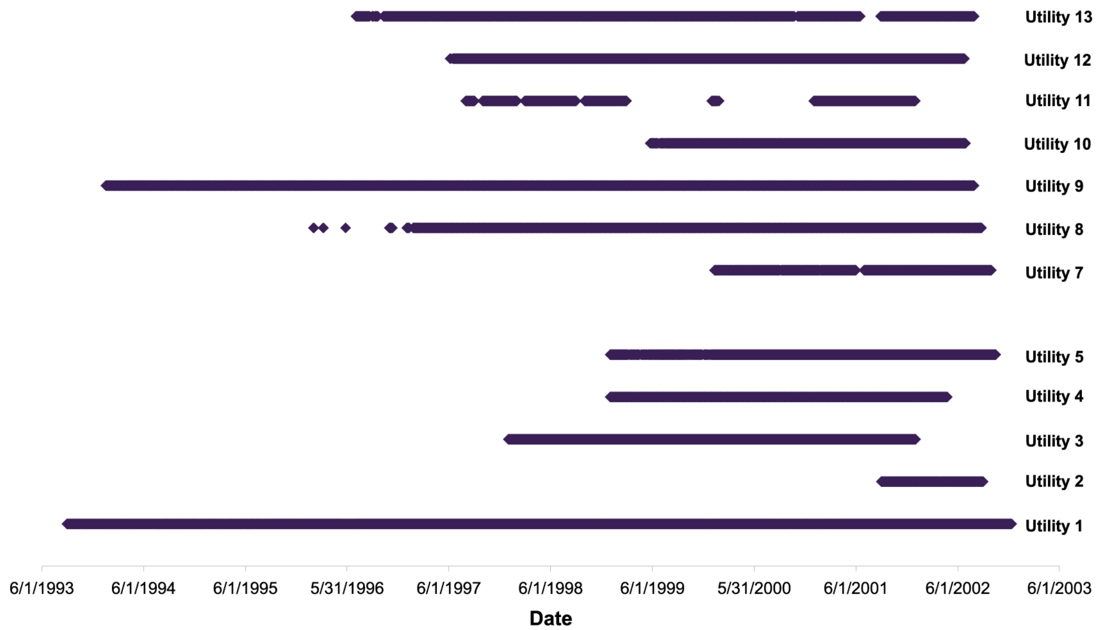



**Figure 1-3**
**DPQ II Monitoring Periods of the 13 Participating Utilities**

**Table 1-1**
**Comparison of DPQ I to DPQ II**

| Characteristic | DPQ I | DPQ II |
|---|---|---|
| Number of sites | 277 | 480 |
| Monitored location categories | 3 | 8 |
| Monitor days | 146,661 | 541,399 |

# 2

# SITE VARIABILITY AND VARIABLES CORRELATED WITH VOLTAGE SAGS

The first step in finding a model to predict voltage sags is to find out what site parameters are likely to affect predictions of voltage sags the most. These are likely to be variables that have the highest correlations to sag rates. This is an issue that's worth investigating on its own right and adds insight on ways to improve power quality to end-use customers. We analyzed the impact of lightning, tree coverage, system voltage, and many circuit parameters on the frequency of voltage sags.

## SARFI Statistics

Several power quality indices have been introduced that are similar to the reliability index SAIFI (System Average Interruption Frequency Index). Utilities can use these for some of the same purposes as reliability indices: targeting areas for maintenance and circuit upgrades, tracking the performance of regions, and documenting performance to regulators. The most widely used power quality index is SARFI, defined as: [1,2]

> $\text{SARFI}_X$, System Average RMS (Variation) Frequency Index: $\text{SARFI}_X$ represents the average number of specified rms variation measurement events that occurred over the assessment period per customer served, where the specified disturbances are those with a magnitude less than *X* for sags or a magnitude greater than *X* for swells.

$$\text{SARFI}_X = \frac{\sum N_i}{N_T} \qquad \textbf{Eq. 2-1}$$

where,

$X$ = rms voltage threshold; possible values - 140, 120, 110, 90, 80, 70, 50, and 10

$N_i$ = number of customers experiencing short-duration voltage deviations with magnitudes above X% for X > 100 or below X% for X < 100 due to measurement event *i*

$N_T$ = number of customers served from the section of the system to be assessed

---

1 *Reliability Benchmarking Application Guide for Utility/Customer PQ Indices,* EPRI, Palo Alto, CA: 1999. 113781.

2 Sabin, D. D., Grebe, T. E., and Sundaram, A., "RMS Voltage Variation Statistical Analysis for a Survey of Distribution System Power Quality Performance," IEEE/PES Winter Meeting Power, February 1999.

The breakpoints were not chosen arbitrarily. The 90, 80, and 70% thresholds are boundaries of the ITI curve, the 50% threshold is a typical breakpoint for motor contactors, and 10% is the dividing line between a sag and an interruption. Two special variations of SARFI have also been defined. $SARFI_{ITIC}$ is the number of events below the lower ITI (Information Technology Industry Council) curve. In similar fashion, $SARFI_{SEMI}$ is the number of events below the SEMI (Semiconductor Equipment and Materials International) curve.

For this analysis, we used a one-minute aggregation for events, and we excluded events with a duration longer than one minute.

The analysis in this report focuses mainly on $SARFI_{ITIC}$, but several other SARFI values are available in the combined dataset, including $SARFI_{SEMI}$, $SARFI_{80}$, $SARFI_{70}$, and $SARFI_{50}$ (see Table 2-1 for statistics on these indices).

**Table 2-1**
**SARFI Site Medians**

| | Transmission Sites (N=53) | Substation Sites (N=183) | Feeder Sites (N=217) |
|---|---|---|---|
| $SARFI_{80}$ | 9.0 | 20.8 | 30.8 |
| $SARFI_{ITIC}$ | 6.3 | 12.6 | 21.5 |
| $SARFI_{70}$ | 5.8 | 10.3 | 20.4 |
| $SARFI_{SEMI}$ | 4.8 | 8.4 | 16.1 |
| $SARFI_{50}$ | 3.3 | 3.8 | 10.9 |

In general, feeder sites had more voltage sags than distribution substation sites, which had more voltage sags than transmission sites. Table 2-2 shows how the site median and upper and lower quartiles changed with location and with survey. Note the wide variation measured at different sites. Figure 2-1 shows this variability with cumulative distributions by monitoring location. With this graph, we see that 80% of distribution feeder sites have a $SARFI_{ITIC}$ of at least 10 events per year. But just 60% of distribution substation sites had $SARFI_{ITIC}$ > 10/year, and only 35% of transmission sites had 10 events below the ITI curve per year.

**Table 2-2**
**$SARFI_{ITIC}$ Site Statistics by Survey and Location**

| Survey | Site Location | P(<25%) | Median P(<50%) | P(<75%) |
|---|---|---|---|---|
| DPQ II | Transmission | 3.5 | 6.3 | 10.7 |
| DPQ II | Substation | 5.7 | 9.4 | 16.1 |
| DPQ II | Feeder | 14.4 | 22.1 | 35.9 |
| DPQ I | Substation | 9.8 | 17.9 | 32.6 |
| DPQ I | Feeder | 11.5 | 21.3 | 36.7 |

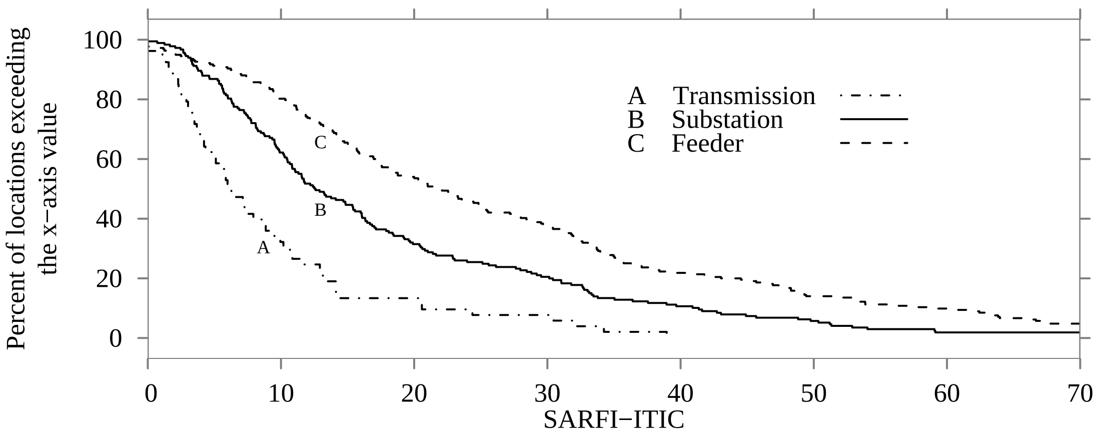



**Figure 2-1**
**Cumulative Distribution of $SARFI_{ITIC}$ by Monitoring Location for the Combined DPQ I/DPQ II Dataset**

In general, we use the site median or actual distributions to characterize or compare the effects of site parameters. As an indicator, the average or mean misrepresents the typical site power quality. The median represents site data better, where by definition, 50% of sites have values higher than the median, and 50% have values lower. With balanced distributions such as the normal distribution, the average equals the median. In a skewed distribution like we see with the power quality data, the average is higher than the median. Additionally, poor sites and anomalies such as a severe storm skew the average upward. In the DPQ data, the average is 31 to 115% higher than the median, depending on which SARFI is used.[3]

The DPQ II study showed much more difference between measurements of sags at substation locations and sags at feeder locations than did the DPQ I study. Note the differences in site medians in Table 2-2. In DPQ I, the feeder sites had a median that was 20% higher than substation sites (21.3 versus 17.9). In DPQ II, the feeder site median was 2.4 times the substation median (22.1 versus 9.4). Much of this difference is due to the difference in where the monitors were located in the two studies. In the DPQ I study, the substation monitors were all downstream of the substation breaker. Whether monitoring the substation bus or the substation feeder, the monitors in the DPQ II were located upstream of the substation breaker. An example of this type

[3] T. A. Short, *Electric Power Distribution Handbook*, CRC Press, Boca Raton, FL, 2004.

of configuration is shown in Figure 2-2. What is seen as an interruption (zero voltage) on the distribution feeder and by monitors in the DPQ I study is seen as just a voltage sag in the DPQ II study. The DPQ II study more accurately shows what happens on the substation bus; the DPQ I study more accurately shows what happens to voltages just outside of the substation.

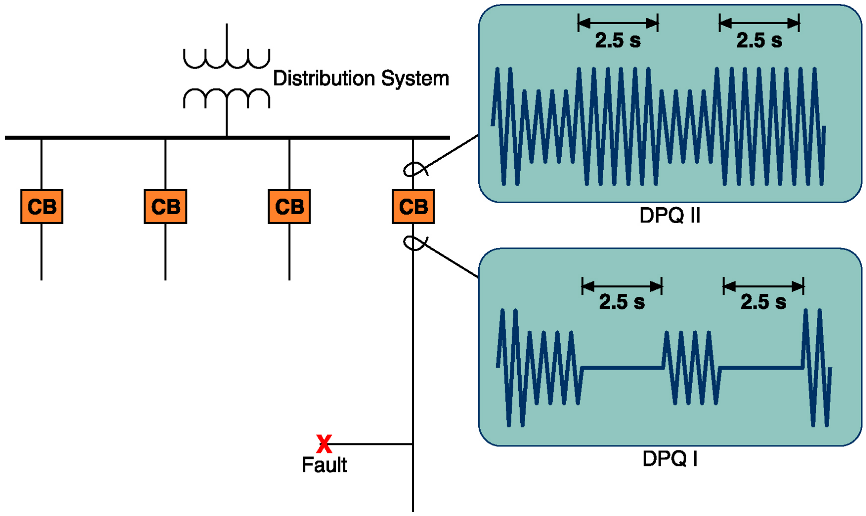



**Figure 2-2**
**Differences in Monitor Connection Between DPQ I and DPQ II for Substation Sites**

## Parameters Available in the DPQ Datasets

A number of site characteristics are available for the DPQ I and DPQ II monitoring locations. Descriptions of these follow:

***Monitoring location –*** Available in both DPQ I and DPQ II datasets. It describes the location of the monitor as one of the following:

- Transmission (DPQ II only)
- Substation
- Feeder

Both the "Substation" and "Feeder" types are on the distribution system. In the DPQ II study, monitoring locations B and E from Figure 1-2 are classified as "Substation," and monitoring locations A and F are classified as "Feeder." In the DPQ I study, locations "Feeder Middle" and "Feeder End" are combined as one classification of "Feeder."

***Load density –*** Available in both DPQ I and DPQ II datasets. Utilities were asked to describe the load density as one of the following:

- Urban
- Suburban
- Rural

- Mixed
- Other

These are mostly based on an estimate by the engineer doing the survey. In the DPQ II dataset, a numeric load density is also available for many sites.

***Load type –*** Available in both DPQ I and DPQ II datasets. Utilities were asked to describe the load density as one of the following:

- Industrial
- Commercial
- Residential
- Mixed

In the DPQ II dataset, the amount of industrial, commercial, and residential load are also available as percentages.

***Feeder construction –*** Available only in the DPQ II dataset. Utilities were asked to describe the feeder construction as one of the following:

- Overhead
- Underground
- Mixed

Where utilities provided circuit lengths, a circuit was classified as overhead if it was more than 2/3 overhead, underground if it was more than 2/3 underground, or mixed.

***Nominal distribution base voltage –*** Available in the DPQ I and DPQ II studies. This is the nominal line-to-line voltage in kilovolts.

***Nominal transmission base voltage –*** Available only in the DPQ II dataset. This is the nominal line-to-line voltage in kilovolts of the subtransmission or transmission system feeding the distribution substation.

***Feeder length –*** Available in both the DPQ I and DPQ II datasets. The question was asked differently during DPQ I than it was asked during DPQ II. In DPQ I, the length is the total circuit miles on just the feeder monitored, broken down as lengths of single-phase and three-phase circuits. In DPQ II, the length is the total circuit exposure in miles of all circuits off the bus. Fortunately, for feeder monitors in the DPQ II study, survey providers appeared to answer with feeder lengths, not total bus lengths. So we used the following:

- Substation sites: Total circuit exposure from all feeders off the substation with length in miles. For DPQ I sites, we multiplied the feeder length provided by the number of feeders on the bus.
- Feeder sites: The circuit exposure in miles on the monitored feeder.

There was ambiguity in the survey form for the DPQ II data, so it is uncertain whether to treat the answers as total three-phase lengths or as three-phase plus single-phase lengths. We analyzed the answers both ways: treating DPQ II lengths as three-phase and treating these lengths as single-phase. For most of the analysis, we assumed that the lengths include all circuits, whether single-phase or three-phase.

***Substation transformer size –*** Available in both the DPQ I and DPQ II datasets. This is the base MVA rating (the open-air or OA rating) of the transformer.

***Number of feeders –*** Available in both the DPQ I and DPQ II datasets. This is the number of feeders off the substation bus. If a utility uses parallel busses (closed bus ties), we divided the number of feeders by the number of busses to obtain a “feeders per bus” number. Therefore, you may notice a non-integer number of feeders in some of the graphs.

***System configuration –*** Available in both the DPQ I and DPQ II datasets. Utilities were asked to describe the system configuration as one of the following:

- Radial
- Primary loop
- Primary selective
- Spot network

***Monitor days –*** Available in both the DPQ I and DPQ II datasets. This is the number of days that the monitor recorded. We sometimes used this to weight distributions or regression models.

***Lightning ground flash density (GFD) –*** Available in both the DPQ I and DPQ II datasets. This is the average annual lightning flashes to the ground per $km^2$ per year. In both the DPQ I and DPQ II studies, latitudes and longitudes of sites were provided. In the DPQ II study, we found latitudes and longitudes from ZIP codes provided by utilities. The GFD is the 10-year average (1988–98) from the U.S. National Lightning Detection Network obtained from EPRI’s Lightning Protection Design Workstation, version 5.0 (see Figure 2-3).

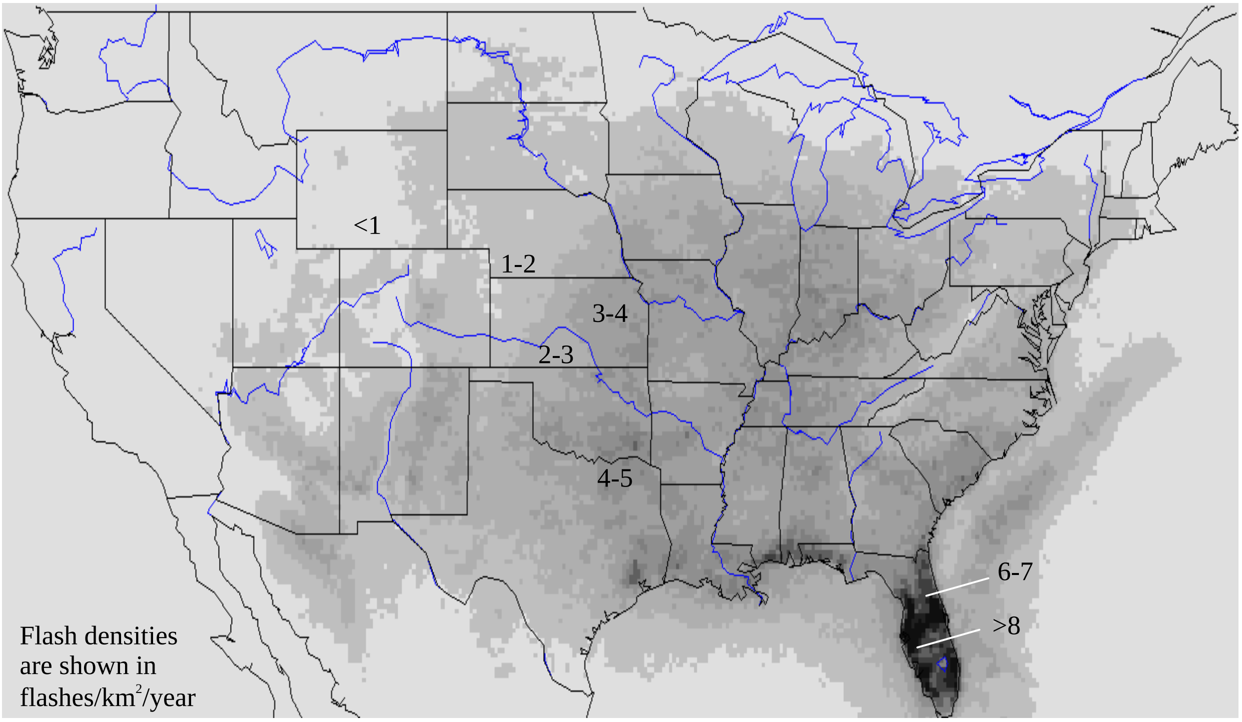



**Figure 2-3**
**Lightning Ground Flash Density Map From the U.S. Lightning Detection Network from EPRI's LPDW v5.0**

***Tree coverage –*** Available in both the DPQ I and DPQ II datasets. This is the average area covered by trees in percent. This is based on the University of Maryland's Global Land Cover Facility data from the Advanced Very High Resolution Radiometer.[4] See Figure 2-4 for an overview. These data are available from:

http://glcf.umiacs.umd.edu/data/treecover/latlongProjection.shtml

Note that this is crude estimate of a line's actual exposure to trees—it is *not* based on any ground-level survey of the circuits in the dataset.

[4] DeFries, R. Hansen, M., Townshend, J.R.G., Janetos, A.C., and Loveland, T.R., *A New Global 1km Data Set of Percent Tree Cover Derived From Remote Sensing*, Global Change Biology, vol. 6, pp. 247–254, 2000.

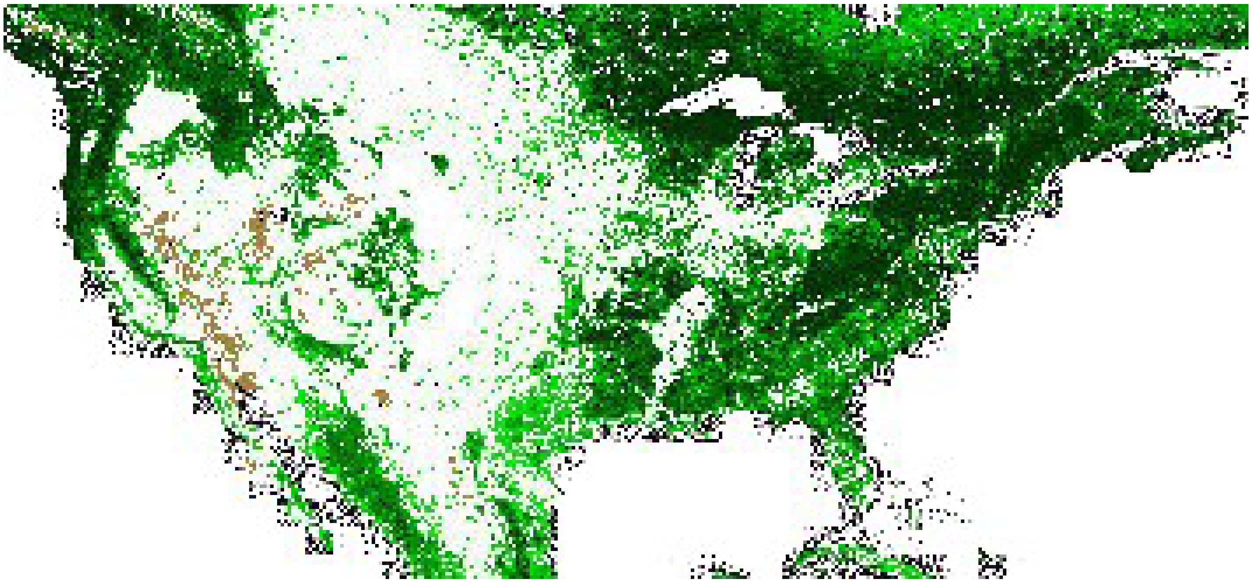
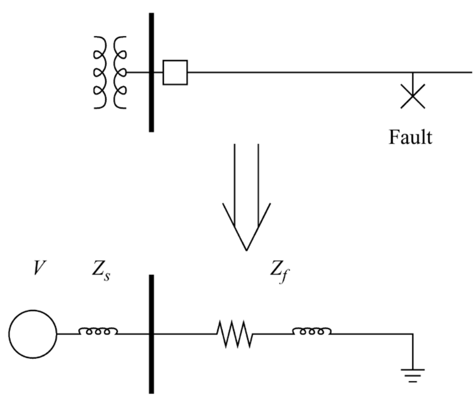


**Figure 2-4**
**Tree-Cover Data for the Continental United States From the University of Maryland's Global Land Cover Facility From the Advanced Very High Resolution Radiometer**

The SARFI statistics at each site from both the DPQ I and DPQ II studies were combined into one dataset. Where possible, site characteristics from the two studies were merged as uniformly as possible. This meant mainly using site characteristics that were common to both datasets. The site characteristics data did have many missing values, which we had to contend with.

## Electrical Parameters That are Important for Sags

The voltage during the fault at the substation bus is given by the voltage-divider expression in Figure 2-5 based on the source impedance ($Z_s$), the feeder line impedance ($Z_f$), and the prefault voltage ($V$). This voltage can be at the substation bus or another location on the power system.

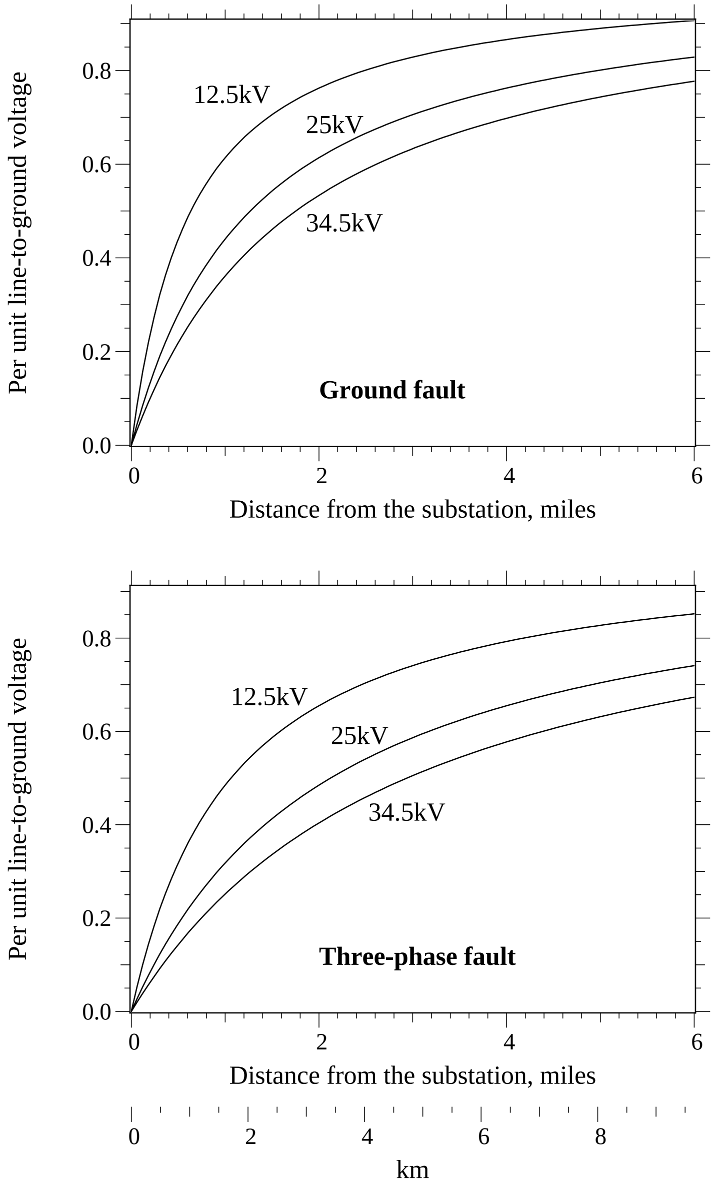



$$V_{bus} = \frac{Z_f}{Z_f + Z_s}$$

**Figure 2-5**
**Voltage Divider Equation Giving the Voltage at the Bus for a Fault Downstream**

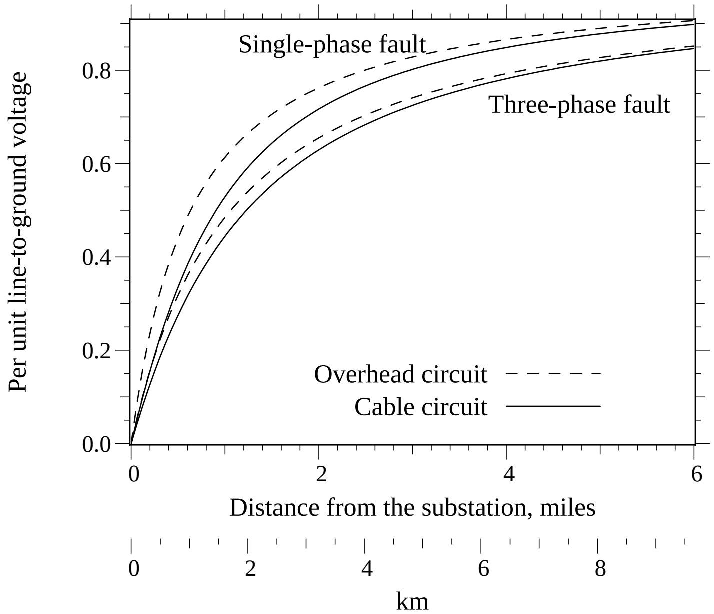
The voltage will sag deeper for faults electrically closer to the bus (smaller $Z_f$). Also, as the available fault current decreases (larger $Z_s$), the sag becomes deeper. The source impedance includes the transformer impedance plus the subtransmission source impedance (often, subtransmission impedance is small enough to be ignored). The impedances used in the equation depend on the type of fault it is. For a three-phase fault (giving the most severe voltage sag), use the positive-sequence impedance ($Z_f = Z_{f1}$). For a line-to-ground fault (the least severe voltage sag), use the loop impedance, which is $Z_f = (2Z_{f1} + Z_{f0})/3$. A good approximation is 1 ohm for the substation transformer (which represents a 7- to 8-kA bus fault current) and 1 ohm per mile (0.6 Ω/km) of overhead line for ground faults. For accuracy, use complex division because the impedances are complex, but for back-of-the-envelope, first-approximation calculations, use the impedance magnitude.

Note that this can be used for any type of fault as long as the appropriate fault values are used in the equation. If the angles are ignored, the equation is an approximation (which is usually acceptable). Figure 2-6 shows a profile of the substation bus voltage for faults at the given distance along the line for 12.47, 24.94, and 34.5 kV. The higher-voltage systems have more severe voltage sags for faults at a given distance. The graph also shows that three-phase faults cause more severe sags. Figure 2-7 compares sags on underground and overhead systems.




Phase characteristics: 500 kcmil, all-aluminum, GMD = 4.69 ft (1.43 m)
Neutral characteristics: 3/0 all-aluminum, 4-ft (1.22-m) line-to-neutral spacing
$Z_1$=0.207 + $j$0.628 Ω/mile (0.1286 + j0.3901 Ω/km)
$Z_0$=0.720 + $j$1.849 Ω/mile (0.4475 + j1.1489 Ω/km)
$Z_s$=0.378 + $j$1.035 Ω/mile (0.2350 + j0.6430 Ω/km)

**Figure 2-6**
**Substation Voltage Profile for Faults at the Given Distance (Single-Phase and Three-Phase Faults Are Shown for Each Voltage)**




500-kcmil aluminum conductor, 220-mil XLPE insulation,
1/3 neutrals, flat spacing, 7.5 in between cables
$Z_1$=0.3543 + $j$0.3596 Ω/mile (0.2201 + j0.2234 Ω/km)
$Z_0$=0.8728 + $j$0.2344 Ω/mile (0.5423 + j0.1456 Ω/km)
$Z_S$=0.5271 + $j$0.3178 Ω/mile (0.3275 + j0.1975 Ω/km)

**Figure 2-7**
**Comparison of Substation Voltage for Faults on Overhead Circuits and Cable Circuits at the Given Distance (Single-Phase and Three-Phase Faults Are Shown)**

The effect of feeder faults on voltage sags at the substation bus can be estimated with the following equation:

$$S(V_{sag}) = n_f \lambda \frac{V_{sag}}{1 - V_{sag}} \left( \frac{Z_s}{Z_f} \right) \quad \textbf{Eq. 2-2}$$

where,

$S$ = annual number of sags per year where the voltage sags below $V_{sag}$

$V_{sag}$ = per-unit voltage sag level of interest (in the range of 0 to 1, such as 0.7)

$n_f$ = number of feeders off the bus

λ = feeder mains fault rate per mile (or other unit of distance) per phase, including faults on laterals and including both temporary and permanent faults

$Z_f$ = feeder impedance in ohms per mile (or other unit of distance); usually use $Z_f$ = $(2Z_1+Z_0)/3$ for ground faults

$Z_s$ = source impedance in ohms

The distribution of voltage sags based on this equation is shown in Figure 2-8 for some common parameters. Several points are noted from this analysis on voltage sags:

- *Exposure* – For 15-kV circuits, we can ignore exposure beyond the first 2 or 3 miles (4 or 5 kilometers) for sags to the bus voltage. The first mile or two is most important as far as circuit improvement, maintenance, or application of current-limiting fuses.
- *System voltage* – Sags are more severe on higher voltage distribution systems (especially at 34.5 kV). A fault 4 miles from the substation sags the voltage much more on a 25-kV system than on a 12-kV system because the substation transformer is a higher impedance relative to the line impedance at higher system voltages. For 24.94 kV, exposure as far as five miles from the station is significant.
- *Single versus three-phase faults* – Three-phase faults cause more severe sags than single-line-to-ground faults. Three-phase faults farther away can pull the voltage down.
- *Underground versus overhead* – All-underground circuits have more exposure to sags because cables have lower impedance than overhead lines.
- *Number of feeders* – The number of sags on the station bus is directly proportional to the number of feeders off the bus.
- *Transformer impedance* – A lower station transformer impedance (a bigger transformer or lower percent impedance) improves voltage sags.
- *Bus tie* – It does not matter whether a substation bus tie is open or closed. If it is open, a fault affects only half of the feeders. A fault that does occur forces a deeper sag because of a higher effective source impedance. These two effects tend to cancel each other.
- *Voltage regulation* – Raising the nominal voltage improves the voltage seen by customers during a fault. Say that a fault drops the voltage to 0.8 per unit, and the prefault voltage was 1.0 per unit. If the prefault voltage were 1.1 per unit, the voltage during the sag is 0.88 per unit. This is not a big difference, but for equipment sensitive to sags to 0.7 to 0.85 per unit, higher voltages appreciably reduce the number of tripouts.

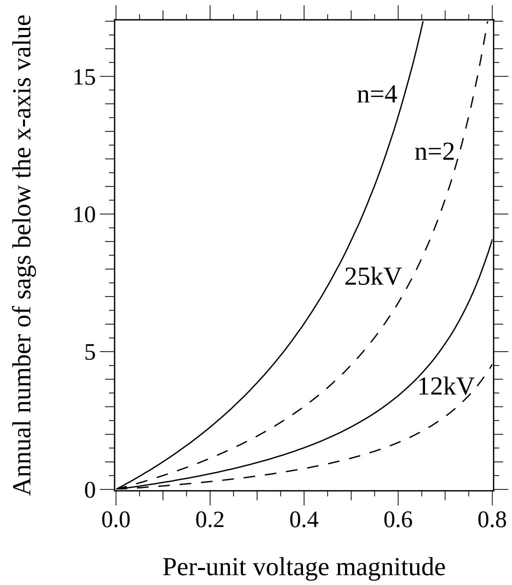



**Figure 2-8**
**Cumulative Distribution of Substation Bus Voltage Sags per Year for the Given (25-MVA, 10% Transformer, 500-kcmil Feeder, n=2 or 4 Feeders off the Bus, λ = 1 Faults/Phase/Mile of Mains/Year, Assumes Line-to-Ground Faults Only)**

Customers at the end of a circuit have more severe voltage sags because almost all faults upstream appear as little or no voltage (most actually fit the power quality definition of an interruption, a voltage to below 10%).

This section explains the theory behind some of the parameters. In the next section, we will investigate the impact of some of these variables on actual data. One of the parameters that we will investigate in more detail is:

$$\frac{n_f \cdot kV^2}{MVA_{xfmr}} \qquad \textbf{Eq. 2-3}$$

The number of bus sags is directly proportional to $n_f$, the number of feeders off the bus, and to $Z_S$, the source impedance (a lower station transformer impedance, a bigger transformer or lower percent impedance, reduces the severity of voltage sags at the station bus). The transformer impedance is $Z_{\%}kV^2/MVA_{xfrmr}$; but because the per-unit impedance of station transformers is roughly constant (7 to 10%), we use $kV^2/MVA_{xfrmr}$.

## Variations With Parameters

Table 2-3 shows relationships between various site characteristics and $SARFI_{ITIC}$ at *feeder* sites. Many trends are expected. Urban sites tend to have fewer sags than suburban or rural sites. Higher-voltage systems have more sags than lower-voltage systems. Circuits in areas with higher lightning have more voltage sags. No one site characteristic dominates. The most pronounced effects are load density, system voltage, and lightning activity.

Keep in mind that the sites have high variability. Figure 2-9 shows much the same data as Table 2-3 but adds bars that range between the lower and upper quartile, so half of the site measurements within the given category fall within the bands shown.

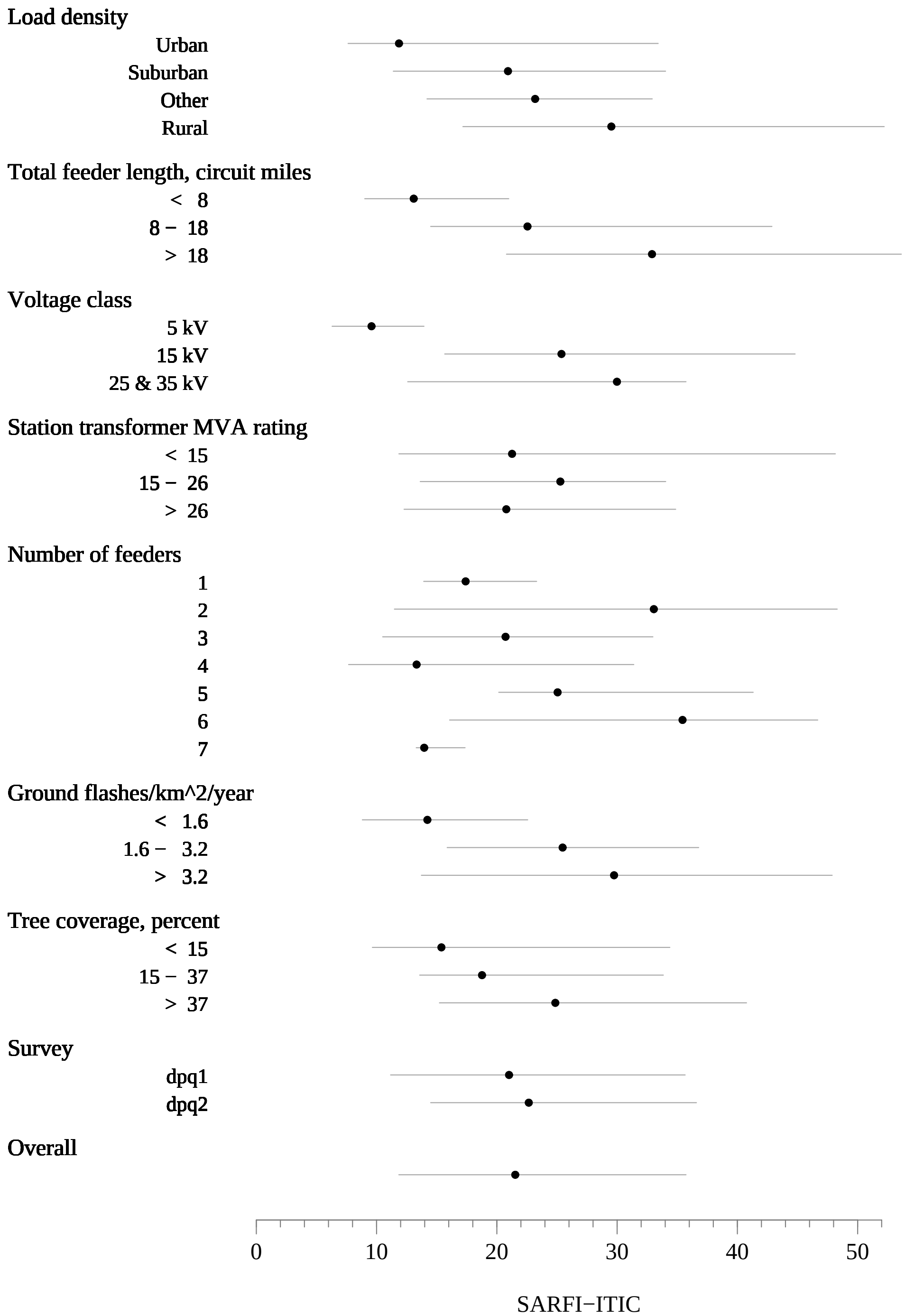
**Table 2-3**
**$SARFI_{ITIC}$ Site Medians by Various Site Characteristics**

**Feeder Sites**

| | $SARFI_{ITIC}$ Site Median | *N* | | $SARFI_{ITIC}$ Site Median | *N* |
|---|---|---|---|---|---|
| Load density | | | Number of feeders | | |
| Urban | 11.9 | 44 | 1 | 17.4 | 31 |
| Suburban | 20.9 | 62 | 2 | 33.0 | 57 |
| Other | 23.2 | 54 | 3 | 20.7 | 48 |
| Rural | 29.5 | 57 | 4 | 13.3 | 22 |
| Total feeder length, circuit miles | | | 5 | 25.1 | 27 |
| < 8 | 13.1 | 48 | 6 | 35.4 | 21 |
| 8 – 18 | 22.5 | 51 | 7 | 14.0 | 5 |
| ≥ 18 | 32.9 | 46 | Ground flashes/km$^2$/year | | |
| Voltage class | | | < 1.6 | 14.2 | 66 |
| 5 kV | 9.6 | 38 | 1.6 – 3.2 | 25.5 | 73 |
| 15 kV | 25.4 | 134 | ≥ 3.2 | 29.8 | 76 |
| 25 & 35 kV | 30.0 | 45 | Tree coverage, percent | | |
| Station transformer MVA rating | | | < 15 | 15.4 | 74 |
| < 15 | 21.3 | 70 | 15 – 37 | 18.8 | 71 |
| 15 – 26 | 25.3 | 71 | ≥ 37 | 24.9 | 70 |
| ≥ 26 | 20.8 | 71 | | | |
| Survey | | | | | |
| DPQ I | 21.0 | 182 | | | |
| DPQ II | 22.7 | 35 | | | |
| | | | All sites | 21.5 | 217 |

*N* = number of sites in the given category. Not all categories have a total of 217 sites because some sites were missing data.

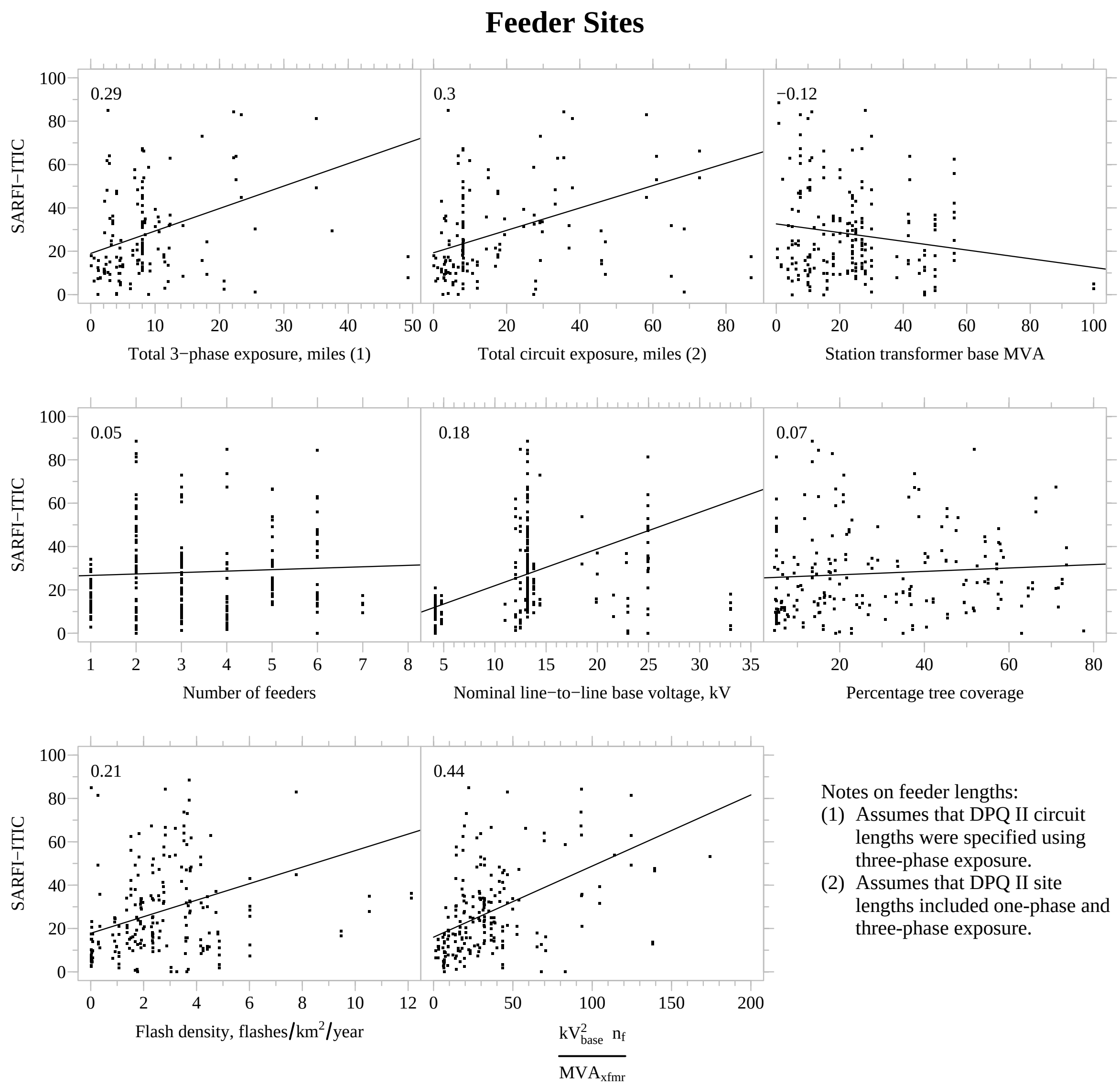
The feeder lengths given in Table 2-3 and Figure 2-9 are circuit miles of the primary, including single- and three-phase circuits. DPQ II feeder site length data were interpreted as including single- and three-phase circuits.




The dots mark the site median, and the bands show the range between the upper and lower quartiles.

**Figure 2-9**
**$SARFI_{ITIC}$ at Feeder Locations Grouped by Several Different Characteristics**

Figure 2-10 shows $SARFI_{ITIC}$ as plotted against various continuous site parameters for feeder sites. The graph also shows a linear curve-fit and a correlation coefficient. Note the high variability of data. Some of the site parameters have some correlation, but again, the data do not show any of the parameters dominating. The terms with the highest correlation are feeder length, system voltage, tree coverage, and the term with transformer impedance and number of feeders ($n_f \times kV^2/MVA_{xfrmr}$).




The correlation coefficients (*r*) are given in the upper-left corner of each plot.

**Figure 2-10**
**$SARFI_{ITIC}$ Site Variations With Various Parameters**

Table 2-4, Figure 2-11, and Figure 2-12 show the effect of site parameters on $SARFI_{ITIC}$ for substation data instead of feeder data. Most of the trends are the same.

Note a new site characteristic available for substation data: The feeder construction parameter (not surprisingly) shows that overhead circuits tend to have more voltage sags than underground circuits. Not enough of the feeder sites had this site characteristic to show a trend for feeders.

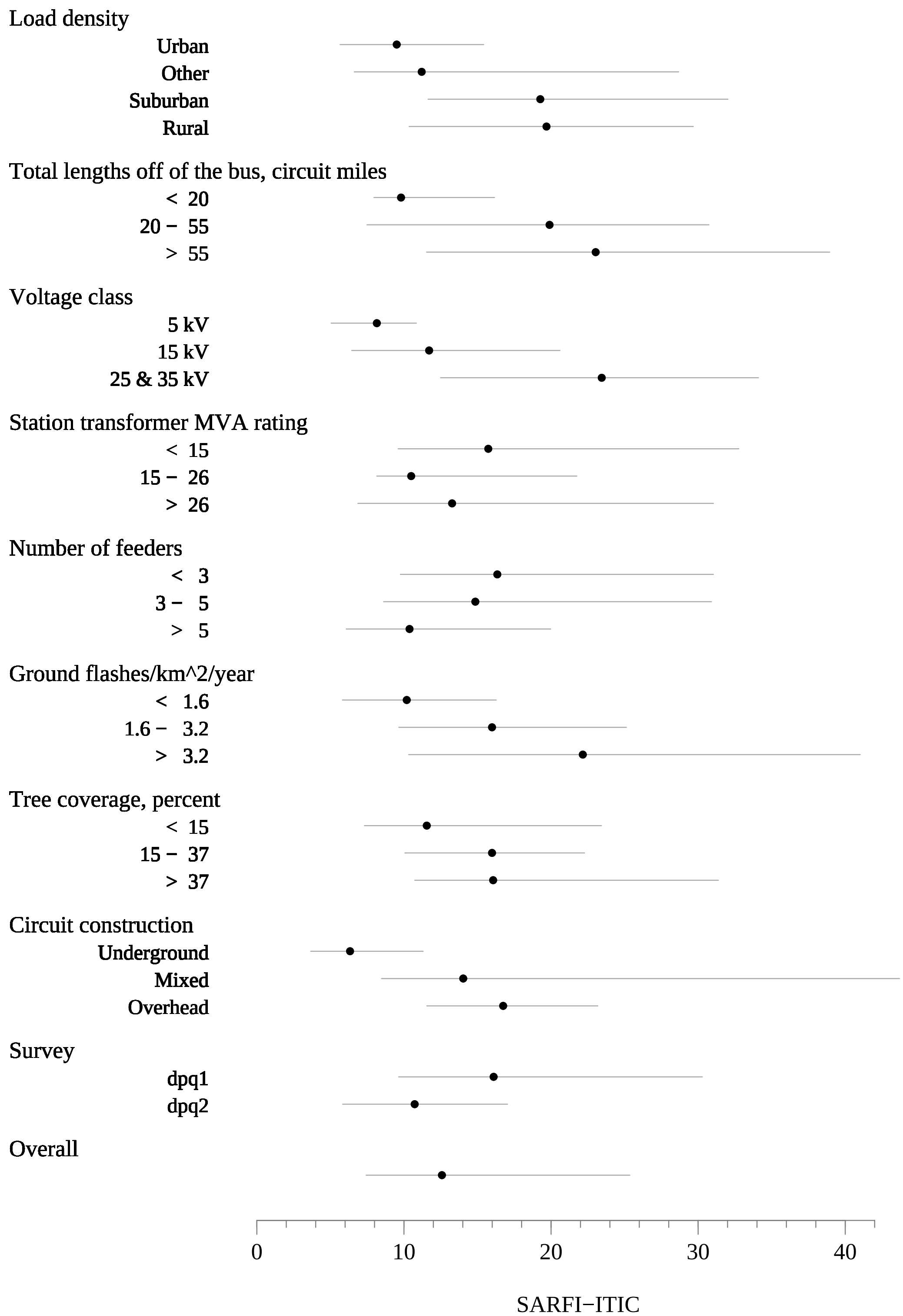
Also, the circuit lengths in these tables and graphs are circuit miles of all feeders fed from the substation bus, whereas the circuit lengths for the feeder sites were just for the feeder where the monitor was.

**Table 2-4**
**$SARFI_{ITIC}$ Site Medians by Various Site Characteristics**

**Substation Sites**

| | $SARFI_{ITIC}$ Site Median | *N* | | $SARFI_{ITIC}$ Site Median | *N* |
|---|---|---|---|---|---|
| Load density | | | Number of feeders | | |
| Urban | 9.5 | 65 | < 3 | 16.3 | 45 |
| Other | 11.2 | 52 | 3 – 5 | 14.9 | 35 |
| Suburban | 19.3 | 31 | ≥ 5 | 10.4 | 26 |
| Rural | 19.7 | 35 | | | |
| Total feeder length, circuit miles | | | Ground flashes/km$^2$/year | | |
| < 20 | 9.8 | 32 | < 1.6 | 10.2 | 52 |
| 20 – 55 | 19.9 | 31 | 1.6 – 3.2 | 16.0 | 69 |
| ≥ 55 | 23.0 | 37 | ≥ 3.2 | 22.2 | 44 |
| Voltage class | | | Tree coverage, percent | | |
| 5 kV | 8.2 | 18 | < 15 | 11.6 | 71 |
| 15 kV | 11.7 | 125 | 15 – 37 | 16.0 | 47 |
| 25 & 35 kV | 23.4 | 40 | ≥ 37 | 16.1 | 47 |
| Station transformer MVA rating | | | Feeder construction | | |
| < 15 | 15.7 | 36 | Underground | 6.3 | 25 |
| 15 – 26 | 10.5 | 35 | Mixed | 14.0 | 20 |
| ≥ 26 | 13.3 | 37 | Overhead | 16.7 | 28 |
| Survey | | | | | |
| DPQ I | 16.1 | 95 | | | |
| DPQ II | 10.7 | 88 | | | |
| | | | All sites | 12.6 | 183 |

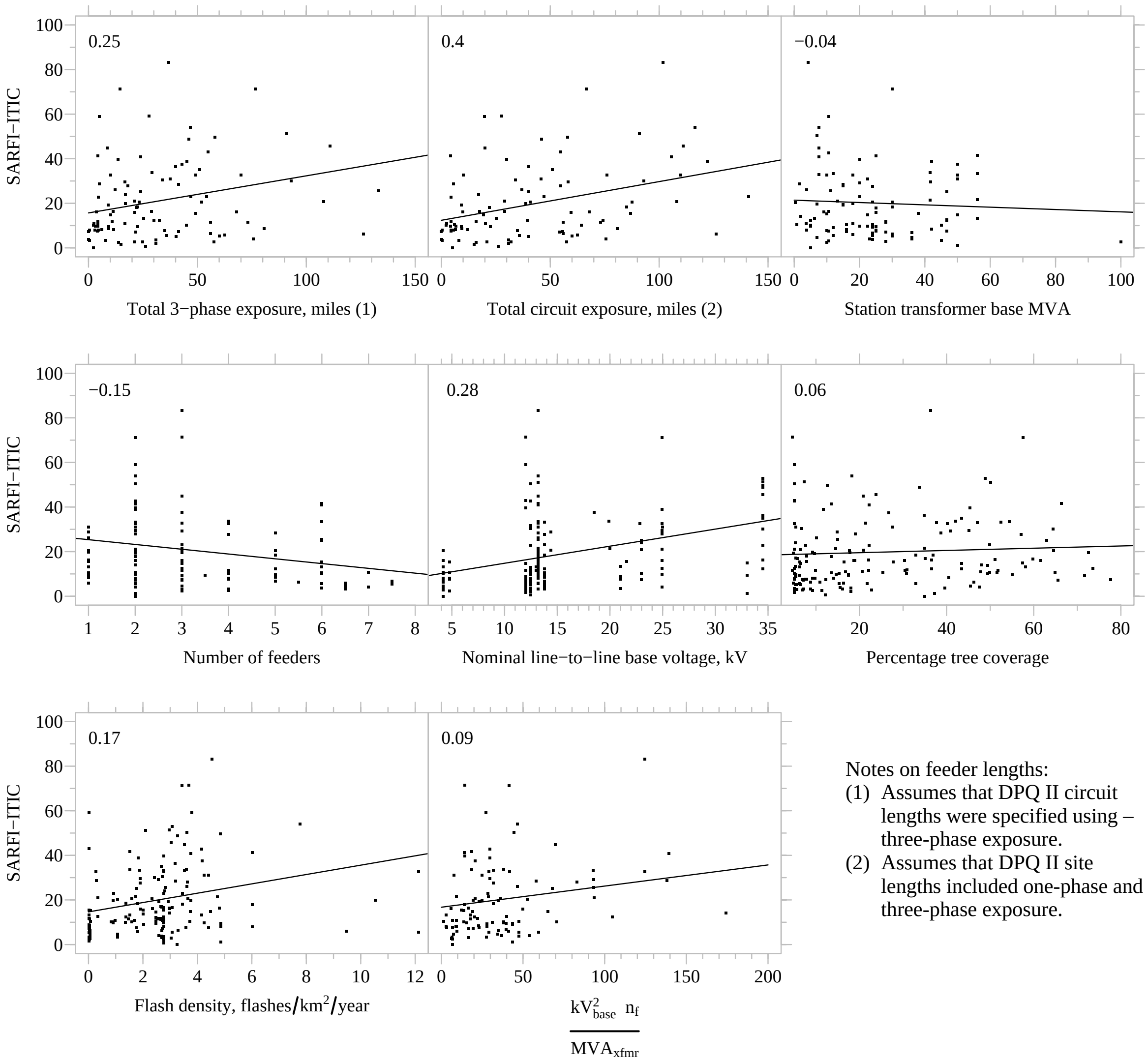
*N* = number of sites in the given category. Not all categories have a total of 183 sites because some sites were missing data.




The dots mark the site median, and the bands show the range between the upper and lower quartiles.

**Figure 2-11**
**SARFI$_{ITIC}$ at Substation Locations Grouped by Several Different Characteristics**

## Substation Sites




The correlation coefficients (*r*) are given in the upper-left corner of each plot.

**Figure 2-12**
**SARFI$_{ITIC}$ Site Variations With Various Parameters**

# 3
# PREDICTION APPROACH

This chapter briefly describes the prediction approach used in this study.

## Model Fitting

Regression techniques are commonly used to find a model prediction formula. Several model types are available. We explored various linear models. For more information on the modeling strategy used here, see Venables and Ripley[1] and McCullagh and Nelder[2].

A linear model is a fit to an equation of the following form:

$$y = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n + \varepsilon \quad \textbf{Eq. 3-1}$$

The $x$'s are site characteristics (such as base voltage or lightning flash density), and the $a$'s are coefficients fitted to the model. The ε is the error distribution, which is Gaussian in the traditional linear model.

A generalized linear model is somewhat different from a standard linear model; we used a generalization where the distribution of the error ε is assumed to be a Gamma distribution rather than a Normal distribution in a strictly linear model. A Gamma distribution skews to the right, like a lognormal distribution. The asymmetric distribution improved the predictions over strictly linear models with a Normal (Gaussian) distribution.

The Gamma error distribution is an important feature of the prediction model. It allows more accurate predictions, and it provides a better estimate of the variability and uncertainty in the answer. A comparison of the distributions of the DPQ I/DPQ II feeder data for $SARFI_{ITIC}$ along with fits to Normal and Gamma distributions are shown in Figure 3-1.

---

1 Venables, W. N. and Ripley, B. D., *Modern Applied Statistics with S-PLUS*. Third Edition, Springer, 1999.
2 McCullagh, P. and Nelder, J., *Generalized Linear Models*, Chapman & Hall, London, U.K., 1989.

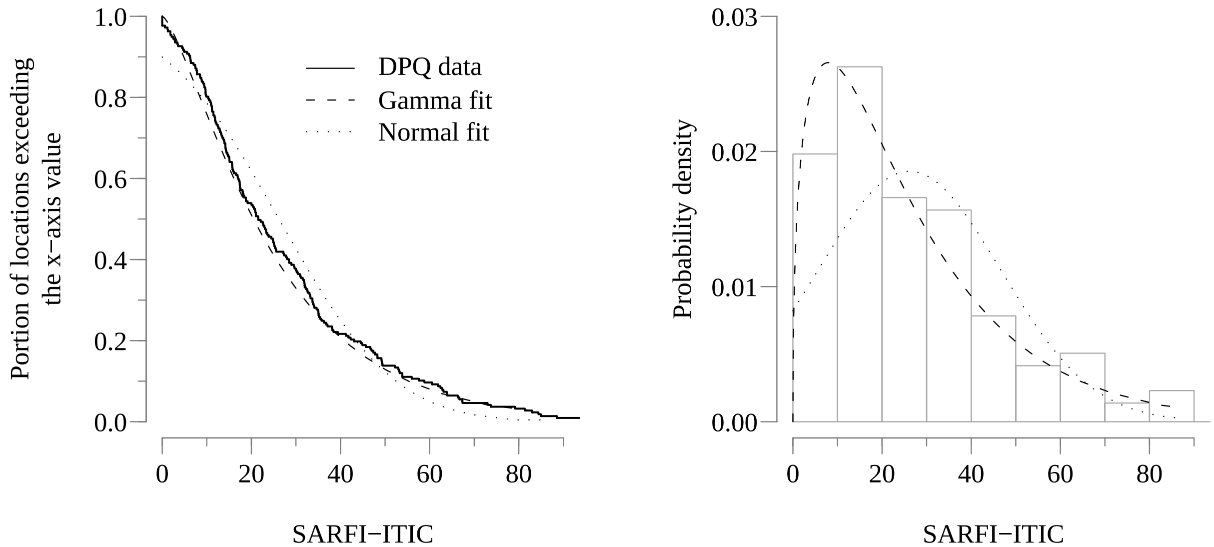



**Figure 3-1**
**Comparison of the SARFI$_{ITIC}$ Data With Normal and Gamma Distribution Fits (Cumulative Distributions and Probability Densities)**

Another common generalized linear model that is often used in reliability modeling is the Gamma distribution with a logarithmic link. This model is of the following form:

$$y = \exp(a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_n x_n) = k_0 \exp(a_1 x_2) \exp(a_1 x_2) \cdots \exp(a_n x_n) \quad \textbf{Eq. 3-2}$$

In this case, the explanatory variables contribute to the prediction multiplicatively rather than additively. We found some logarithmic-link models that were almost as good as the additive linear model with the Gamma distribution. But because they were no better than strictly linear models, the linear models were used.

Several different methods are available for choosing models. The most straightforward is the deviance, the measure that is being optimized. For a standard linear model, the deviance is the residual sum of squares. For a generalized linear model, the deviance is the log-likelihood. Another more general way to grade models is with the AIC (Akaike Information Criterion). The AIC includes a penalty on the number of terms as well as the log likelihood (adding more terms to a model reduces the degrees of freedom). We chose models with the lowest AIC score.

Most of the modeling on this project was developed using the *S* language,[3] either the open-source version known as *R* (www.r-project.org) and/or the commercial S-PLUS software (www.insightful.com).

## Missing Data

Many site characteristics in the combined DPQ I/DPQ II dataset are unknown; they were not filled in by utilities during the site-characterization stage of the project. Figure 3-2 shows a graph

[3] Becker, R. A., Chambers, J. M., and Wilks, A. R., *The NEW S Language*, Chapman & Hall, 1988.

highlighting the missing data for several important site characteristics. When we have data missing, the traditional options normally include:

- Remove sites with incomplete data.
- Remove variables with missing values.

Neither of these options was particularly attractive with our combined power quality dataset—very few sites have complete site characteristic data, and almost all important variables have sites with unknown data. Even a combination of site and variable removal would leave a very reduced dataset.

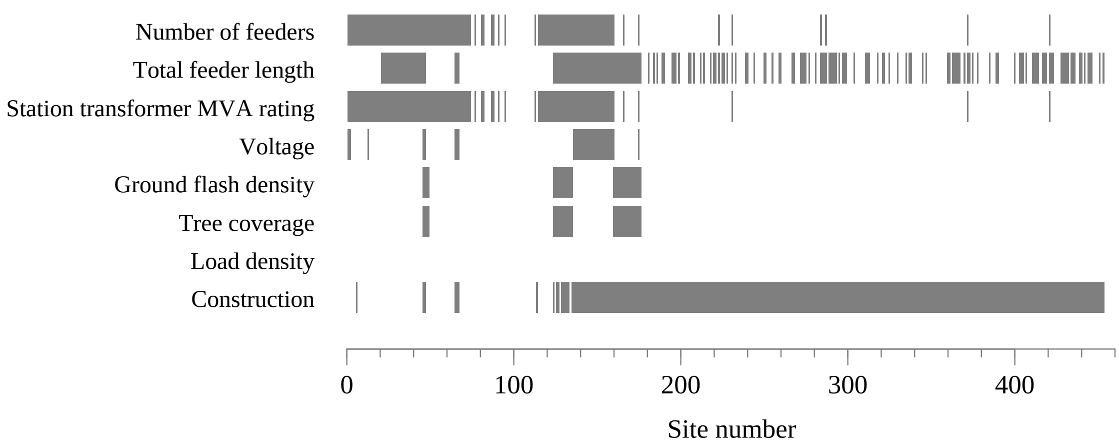



**Figure 3-2**
**Missing Site Data in the Combined DPQ I/DPQ II Dataset**

Fortunately, there's a way to fill in missing data, so we can fit models without distorting the results. The approach is called *multiple imputation*.[4,5] Consider a method called single imputation, where we impute (fill in) missing data with approximate answers, say fill in all missing data with the average value of that variable. Then we could analyze the dataset as if it were a complete dataset. However, problems arise because by filling in missing values with that variable's average, we distort the effect of that variable—that variable has less variability than it should, and this lack of variability distorts the effect of that variable on the prediction model.

To more accurately fill in data requires multiple imputations—we basically make several copies of the original dataset and fill in each dataset with imputations drawn (randomly or by some other means). The main steps in multiple imputation are:

1. *Multiple imputations* – Copy the original dataset to several copies. With each copy, fill in missing values with imputations. Choose imputations such that they reflect the variability and correlations with other parameters. Surprisingly, only three to 10 imputed datasets are needed to adequately reflect the variability and correlations in the missing data.

[4] Little, R. J. A. and Rubin, D. B., *Statistical Analysis With Missing Data*, 2nd Edition, J. Wiley & Sons, New York, 2002.
[5] Schafer, J. L., *Analysis of Incomplete Multivariate Data*, Chapman & Hall, London, 1997.

2. *Analysis* – Analyze each imputed copy of the original dataset as if it is a complete dataset. This includes finding a regression model for predicting a result.

3. *Pooling* – Combine results from the analysis (model coefficients or predictions) by averaging or taking the median.

Step 1 is the difficult step. Variables with missing data are not always independent; they can be correlated with other (dependent or independent) variables (some of which may also have missing data). The fact that data are missing may be a function of the value of the variable (or other variables).

The software we used for multiple imputation is the *MICE* library by S. van Buuren and C. G. M. Oudshoorn, available for the R and S-PLUS systems (http://www.multiple-imputation.com/).

Multiple imputation can be used to fill in missing values in the original dataset. This is the easiest approach because multiple imputations only need to be done originally. The resulting prediction model is simple enough to include in a spreadsheet—no further imputation-related manipulation or analysis is needed. Multiple imputation can also be used to fill in missing values at a new site where some information is unknown, but we still need a prediction. This is more complicated for the user, because the user needs to employ rather sophisticated statistical software to get results (it is now more complicated than a spreadsheet can handle).

## Predicting Variability

One of the most important results of a prediction is an estimate of the variability and accuracy of the model. Two measures are often used to quantify the precision:

- *Confidence interval* – Of all sites with the given site characteristics, the average SARFI of all sites will fill within this interval at a given percentage of confidence (say 95%).
- *Prediction interval* – Of all sites with the given site characteristics, a certain percentage (again say 95%) of them will fall within the interval. Or said another way, there is a given probability that a specific site will fall within the prediction interval.

For the same percent probability, the prediction interval is wider than the confidence interval. For predictions, the prediction interval is much more useful and is used in this project.

For a linear model with normal error distribution of the form:

$$y = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n + \varepsilon \qquad \textbf{Eq. 3-3}$$

the prediction interval calculation can be directly solved. For a generalized linear model, the prediction intervals are more difficult to find. As a first approximation, the prediction intervals may be found from the distribution of the error term, which we have assumed to have a Gamma distribution. We refer to these as “simple prediction limits” in this report. The problem with these simple prediction limits is that they do not include any variability in the prediction due to uncertainty in the model parameters ($a_1$, $a_2$, …).

One way to estimate prediction intervals is with a method called the *bootstrap*.[6] To bootstrap, take a random sampling of data with replacement. The resampling *with replacement* is the key to the bootstrap—by randomly taking a set of values from the original data, we get slightly different versions of the data; note that some of the data will be repeated because of sampling with replacement. If we take enough bootstrap copies of the dataset (say 1000) and perform some analysis of each of the 1000 datasets, then the variability of the answers (whether it is a prediction or some other analysis) gives us an estimate of the variability of the answer relative to the entire population. Bootstrapping assumes that the original data are a good representation of the overall distribution. Bootstrapping does not make any assumption about the underlying distribution. It is also very flexible and can be adapted to many types of problems. A major downside to bootstrapping is that it requires specialized software, and even with modern computers, the processing time for each run can be several minutes.

For prediction interval estimation, the steps to finding prediction intervals are:[7]

1. Find 1000 bootstrap copies of the data.

2. For each copy of the bootstrap data, fit a generalized linear model to the data.

3. For each model, estimate a prediction from that model using the site characteristics of interest.

4. Add a random error to the prediction from the Gamma error distribution ($\varepsilon$) for that model.

5. Sort the 1000 predictions. This gives a distribution of possible predictions that includes errors due to the coefficients and errors due to the error term.

6. Find prediction intervals as appropriate. For example, with 1000 bootstrap predictions, the 950th term in the sorted prediction list represents the prediction threshold such that 95% of sites should have a value less than that.

The software we used for bootstrapping is the Angelo Canty's *BOOT* library, which is available for R and S-PLUS. For the R version, see http://cran.r-project.org, and for the S-PLUS version, see http://statwww.epfl.ch/davison/BMA/library.html.

It is possible to include multiple imputations to handle missing data within a bootstrap approach. For each bootstrap, also replace missing values with imputations from the set of already estimated multiple imputations; as a first approximation, you do not have to redo the multiple imputations at each step.

As it turns out, the bootstrap approach to prediction levels is not vital for predicting voltage sags. The Gamma error distribution ($\varepsilon$) dominates, so it is possible to directly use the Gamma distribution to estimate prediction intervals of SARFI predictions. This makes it possible to

6 Davison, A. C. and Hinkley, D. V., Bootstrap Methods and Their Applications, Cambridge University Press, Cambridge, 1997.

7 There is more than one way to bootstrap prediction intervals. This is the way we feel is most appropriate for the prediction models that we developed. For more information, see Davison and Hinkley, 1997.

produce useful estimates of variability in a spreadsheet-type model; advanced statistical software is not needed. Where the bootstrap approach is necessary is incorporating measurement data as discussed in the next section.

## Incorporating Measurement Data

If we have some power quality monitoring at a site, it would be nice to use that data to improve our prediction about voltage sags. As Bollen's work shows in Table 3-1, it can take many years for accurate estimates of SARFI at a site. But with monitoring coupled with a site prediction model, we may improve the estimate.

**Table 3-1**
**Monitoring Duration Needed to Achieve the Given Level of Accuracy**

| Event frequency | 50% accuracy | 10% accuracy |
|---|---|---|
| 1 per day | 2 weeks | 1 year |
| 1 per week | 4 months | 7 years |
| 1 per month | 1 year | 30 years |
| 1 per year | 16 years | 400 years |

Source: Bollen, M. H. J., *Understanding Power Quality Problems: Voltage Sags and Interruptions*, IEEE Press, New York, 2000.

There is a way to use bootstrapped prediction models along with measured data to estimate a new prediction. The approach relies on Bayesian statistics, where probability of a new event can be calculated based on an earlier probability estimate.

The basic idea is to use a bootstrap of model coefficients as a prior distribution, then to use the fundamental Bayesian formula to obtain a new distribution, which will be in the form of a weighted distribution. The way this works is as follows:

1. Bootstrap the prediction models as described in the previous section (1000 cases).

2. Each of the 1000 bootstraps produces a prediction model. Each has a set of coefficients ($a_1$, $a_2$, …) and an error term (ε).

3. Using a given site's characteristics, find the prediction for each of the 1000 bootstraps.

4. For $y_1$, a measurement of $SARFI_X$ from one year of monitoring, find the probability of that measurement occurring for each of the 1000 bootstraps. Find this probability using the prediction from step 3 and the dispersion of the Gamma distribution. Each of these 1000 probabilities represents a weighting of the 1000 error distributions.

5. Repeat for each year of monitoring ($y_2$, $y_3$, …).

6. Combine the results by multiplying successive probabilities at each site. This means the probability weight for bootstrap number one is $w_1 = P_1(y_1) \times P_1(y_2) \times P_1(y_3) \times \ldots$ , and this is repeated for each of the 1000 bootstraps.

7. Now each of the 1000 weights represents a weighting on that prediction model. A new prediction estimate may be obtained from a weighted average. A new distribution of the results may be shown with a weighted probability distribution plot or prediction intervals may be obtained from a weighted quantile evaluation.

In this approach, we have assumed that each SARFI prediction represents one year of measurement at a given site. Based on the DPQ data, this assumption is not quite valid, as we will see in the next section, but it is close.

This method is complicated, but an advantage is that we do not have to repeat the bootstrapping for every site that we want to evaluate. We only need to bootstrap on the original data (as long as we are happy with the simple prediction limits rather than the bootstrapped prediction limits). That said, it is still complicated enough such that it would be difficult to implement this methodology into a spreadsheet, making it easier for a user.

# 4 PREDICTIONS FOR SARFI

Now, we get to the heart of the matter—predictions. In this section, prediction models are developed using DPQ Phase I and Phase II data. We also discuss extensions that can incorporate measurement data into a prediction. The variability in the data and the variability and accuracy of the predictions are also discussed.

For this work, we concentrate on predictions of voltage sags at *feeder* sites using a combined dataset from the DPQ Phase I and DPQ Phase II studies. Using feeder sites is more representative of the voltage to customers. Also, using feeder sites avoids a significant difference in measurements between DPQ Phase I sites and DPQ Phase II sites (DPQ Phase II sites measure substation bus voltage, so they exclude many momentary interruptions).

## Predicting Sags Based on Predictions Derived From DPQ Phase I Data

The starting point for predictions is a formula for $SARFI_{ITIC}$ based on DPQ Phase I data. It is a linear equation as follows:

$$N_{ITIC} = 4.74 + 0.293l + 2.47N_g + 0.192\frac{n_f \cdot kV^2}{MVA_{xfmr}} + 8.2 \text{ if moderate to heavy tree coverage} \quad \textbf{Eq. 4-1}$$

where,

$N_{ITIC}$ = predicted annual number of events that fall under the lower ITI curve

$l$ = total exposure (including three-phase and single-phase portions) on the circuit in miles (multiply kilometers by 1.609)

$N_g$ = lightning ground flash density in flashes/km$^2$/year

$kV$ = base line-to-line voltage in kV

$n_f$ = total number of feeders off the substation bus

$MVA_{xfmr}$ = station transformer base rating (open-air rating) in MVA

See Appendix A for more information on this model. The DPQ Phase I model works well for DPQ Phase II data. Figure 4-1 shows predicted values versus site observations for the DPQ II data. Figure 4-2 shows the same type of graph, but with the original DPQ I data that was used to derive the model; about the same amount of variation exists. Table 4-1 shows how many of the site predictions fall within various prediction bands. The DPQ I model holds well for both the DPQ I and the DPQ II data.

Note the high degree of variability in the predictions—these are not precise estimates, but they do well in predicting observations within a given prediction band.

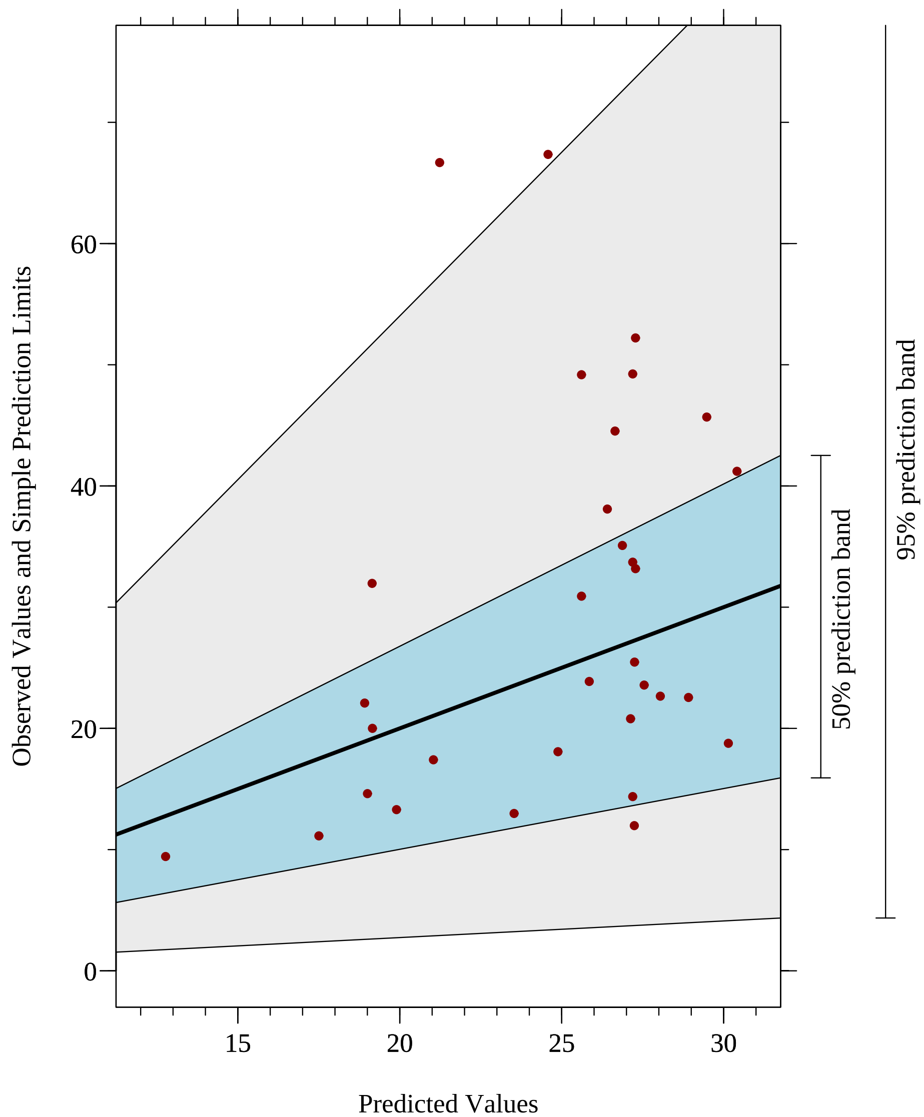



**Figure 4-1**
**Observed $SARFI_{ITIC}$ Versus Predicted $SARFI_{ITIC}$ for DPQ II Feeder Site Data Using the DPQ I Model**[1]

[1] T. A. Short, A. Mansoor, W. Sunderman, and A. Sundaram, "Site Variation and Prediction of Power Quality," *IEEE Transactions on Power Delivery*, vol. 18, no. 4, pp. 1369–75, October 2003.

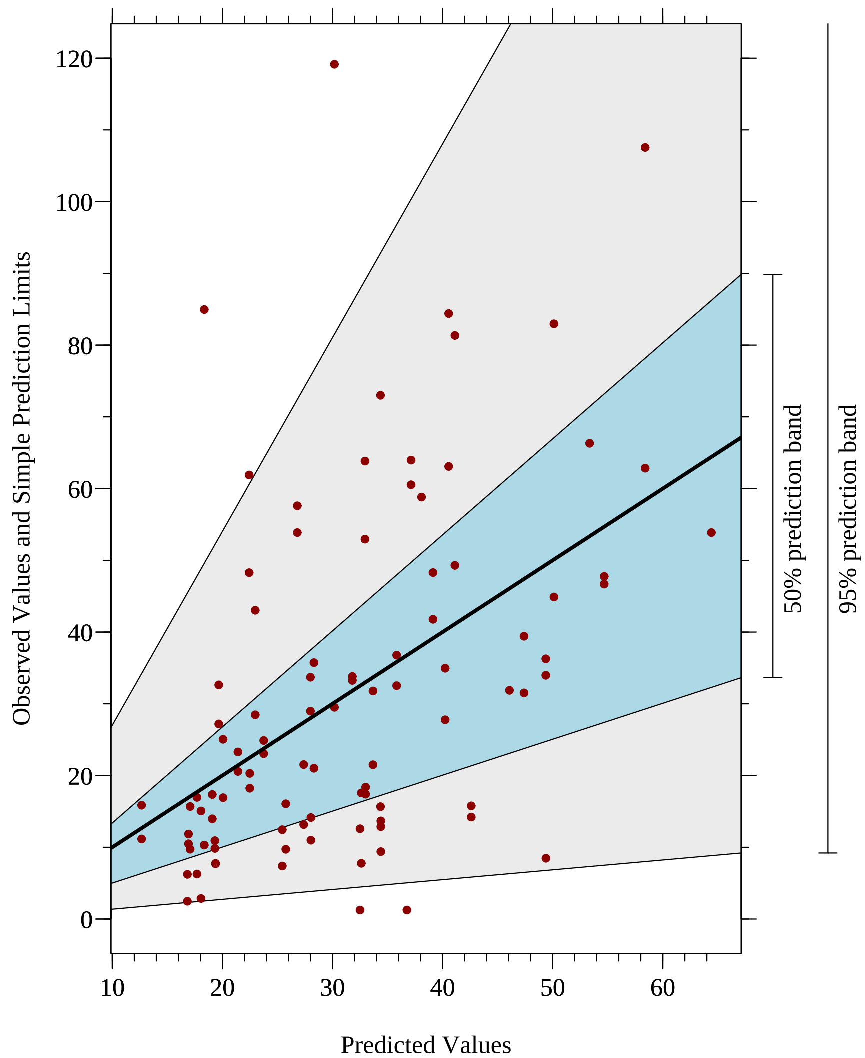



**Figure 4-2**
**Observed $SARFI_{ITIC}$ Versus Predicted $SARFI_{ITIC}$ for DPQ I Feeder Site Data Using the DPQ I Model**

**Table 4-1**
**Percent of Sites That Have a $SARFI_{ITIC}$ That Falls Within the Given Prediction Band for $SARFI_{ITIC}$**

| Data source | | Prediction band in percent | | | |
|---|---|---|---|---|---|
| | | 25% | 50% | 75% | 95% |
| DPQ I data — | the dataset used to derive the prediction model | 33% | 57% | 80% | 95% |
| DPQ II data — | new data | 34% | 65% | 84% | 94% |

## Updated SARFI$_{ITIC}$ Predictions

By adding the DPQ II data and using the multiple imputation methodology, which allows evaluation of missing data, the feeder dataset increases from 105 to 217 sites, and the monitoring years increase from 195 to 366 years. This combined dataset yielded similar results to the original DPQ I prediction analysis. The new equation for predicting SARFI$_{ITIC}$ is:

$$N_{ITIC} = 4.26 + 0.46l + 2.2N_g + 0.247\frac{n_f \cdot kV^2}{MVA_{xfmr}} + 6.33 \text{ if moderate to heavy tree coverage} \qquad \textbf{Eq. 4-2}$$

where,

$N_{ITIC}$ = predicted annual number of events which fall under the lower ITI curve

$l$ = total exposure (including three-phase and single-phase portions) on the circuit in miles (multiply kilometers by 1.609)

$N_g$ = lightning ground flash density in flashes/km$^2$/year

$kV$ = base line-to-line voltage in kV

$n_f$ = total number of feeders off of the substation bus

$MVA_{xfmr}$ = station transformer base rating (open-air rating) in MVA

Several variations of this model were attempted, and variations of logarithmic models were also tried, but this form held as the optimal fit.

## Variability and Uncertainty

One must keep in mind that these are not precise predictions. Rather than give a precise prediction, the predictions narrow the probability range for SARFI at a given site. One of the outputs of the regression model is an estimate of the dispersion.

The error term on the regression model fits a Gamma distribution with a given *dispersion* parameter. For the SARFI$_{ITIC}$ prediction formula, the dispersion = 0.389. A Gamma distribution has two parameters: the shape (often denoted by $\alpha$) and the scale (often denoted by $\beta$). The mean of a Gamma distribution is $\alpha\times\beta$. The GLM regression finds an error term with a dispersion = $1/\alpha$. Therefore, the shape and scale are related to the prediction (the mean) and dispersion parameter by:

$$\alpha = 1/\text{dispersion}$$
$$\beta = \text{prediction} \times \text{dispersion} \qquad \textbf{Eq. 4-3}$$

In Microsoft Excel, the formulation to find the value of SARFI exceeded by a given percentage of sites is:

$$= GAMMAINV((100 - prob)/100, 1/dispersion, pred * dispersion) \qquad \textbf{Eq. 4-4}$$

where

prob = probable number of sites in percent that will have a SARFI exceeding the result

dispersion = dispersion parameter of the gamma distribution

pred = prediction for a given site based in its site parameters

Figure 4-3 shows the probability densities of two predictions for $SARFI_{ITIC}$. The prediction of 19 (the left curve) has a tighter predicted distribution than does the prediction of 55. Note that with the simple prediction distributions, all sites that have a prediction of $SARFI_{ITIC}$=19 have the same error distribution, regardless of the differences in input parameters (this is not true with the more rigorous bootstrap prediction distributions).

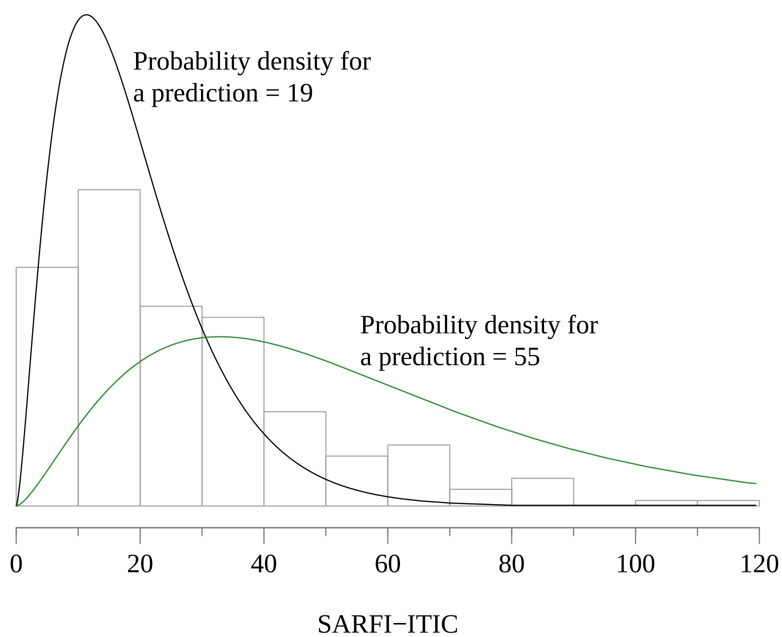



**Figure 4-3**
**Probability Density Functions of the Gamma Distribution for Two Different Predictions (dispersion=0.389) and Histogram of the DPQ I/DPQ II Feeder Data**

The variability in the model prediction includes three main components:

- *Time variability* – Most of the site data have a rather short monitoring period. The significant year-to-year variations increase the uncertainty in the model.
- *Site variability* – Voltage-sag rates vary widely from site to site.
- *Model uncertainty* – Many more factors affect voltage sags than the simple models used here.

It is difficult to separate these effects from each other with the DPQ dataset.

## Other Prediction Formulas

While this analysis has mainly concentrated on $SARFI_{ITIC}$, other voltage sag predictions are possible. Each of these predictions fits the following form, where the coefficients for the given index are in Table 4-2:

$$SARFI_X = k_0 + k_1 l + k_2 N_g + k_3 \frac{n_f \cdot kV^2}{MVA_{xfmr}} + k_4 \text{ if moderate to heavy tree coverage} \quad \textbf{Eq. 4-5}$$

**Table 4-2**
**Coefficients for Prediction Equations for Various SARFI Indices**

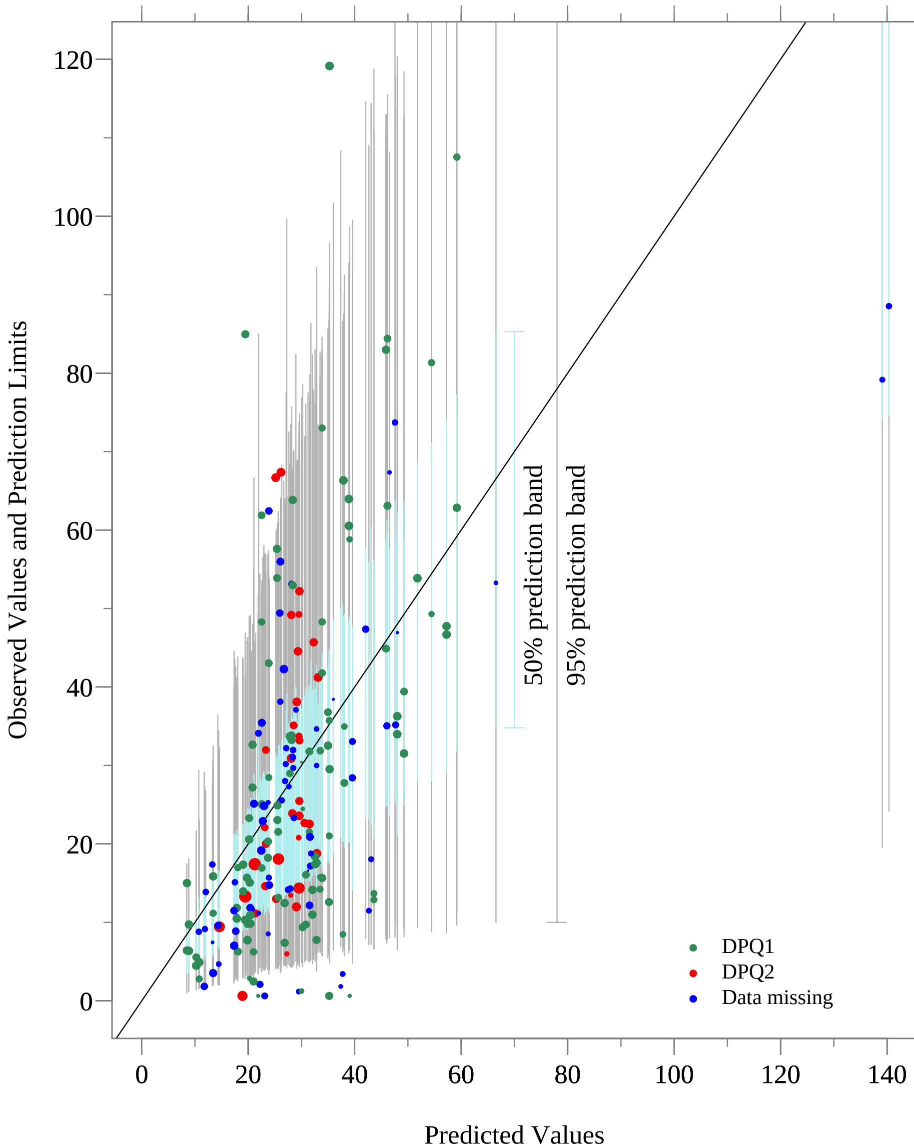
| | $k_0$ Intercept | $k_1$ $l$ | $k_2$ $N_g$ | $k_3$ $\frac{n_f \cdot kV^2}{MVA_{xfmr}}$ | $k_4$ Tree term | Dispersion |
|---|---|---|---|---|---|---|
| $SARFI_{80}$ | 11.99 | 0.426 | 2.71 | 0.322 | 3.55 | 0.399 |
| $SARFI_{ITIC}$ | 4.26 | 0.480 | 2.20 | 0.247 | 6.33 | 0.389 |
| $SARFI_{70}$ | 6.23 | 0.353 | 2.24 | 0.195 | 4.37 | 0.435 |
| $SARFI_{SEMI}$ | 3.50 | 0.356 | 1.56 | 0.135 | 6.18 | 0.436 |
| $SARFI_{50}$ | 4.33 | 0.358 | 1.64 | 0.000 | 3.89 | 0.552 |

We did some evaluations of different models (dropping terms, trying different error models, or trying other regression types), and the general structure of the $SARFI_{ITIC}$ model holds well for the other SARFI indices. But note that not as much care has gone into the model selection forms and variables, so although the other SARFI models should hold well, they are not necessarily optimal.

## Extensions to the Prediction Models

Using bootstrapping to refine our estimate of prediction intervals was explored. Bootstrapping gives a better estimate of prediction bands. The simple prediction limits shown in the previous sections did not include all of the variation possible. One set of site parameters may produce more variation than another set of site parameters. Bootstrapping is a way to include the variability of the model. Figure 4-4 compares predictions versus observations along with prediction bands. Figure 4-4 is similar to Figure 4-1and Figure 4-2, but because we have bootstrapped each prediction, each prediction has its own prediction limit, so these are drawn as individual bands associated with each prediction. By changing the color and size of the observed data points, the figure also allows us to explore the effects of monitoring duration, survey, and missing data on the outliers (a number of which have missing data or short monitoring periods).

The main result drawn from Figure 4-4 is that the bootstrapped prediction intervals are not sufficiently more accurate than the simple prediction intervals. The simple prediction intervals are sufficient because the error term dominates the error from model variations. This is welcome news because it allows a simpler implementation for users.




Note: The area of the dots is proportional to the monitoring period at the given site.

**Figure 4-4**
**Observed $SARFI_{ITIC}$ Versus Predicted $SARFI_{ITIC}$ for Feeder Site Data Using the Bootstrap Prediction Bands**

## Predictions Based on Yearly Data

When interpreting the SARFI predictions and the variability, it is important to know what the result represents. The estimate of SARFI is supposed to be the long-term average for the site, but it may be closer to a short-term estimate because the DPQ data generally have a relatively short monitoring period. As an exercise, we split the data into calendar years and derived predictions based on this yearly data. Figure 4-5 compares cumulative distributions of site average and yearly data. Both have a high degree of variability. As expected, the yearly values have more variability and more zeros in the data.

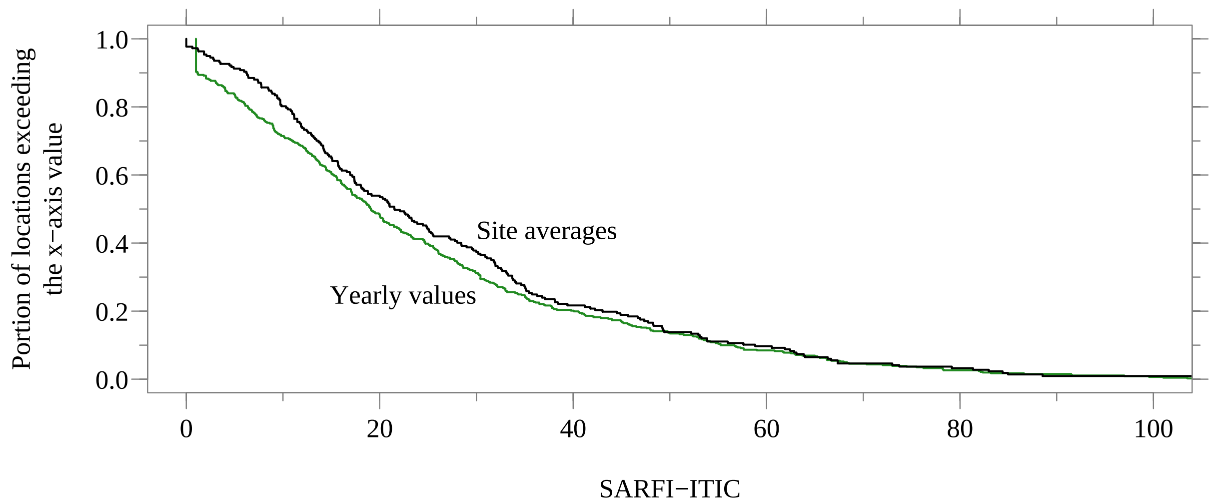



**Figure 4-5**
**Cumulative Distributions of $SARFI_{ITIC}$ Based on Site Averages and on Yearly Values**

Table 4-3 compares a prediction model based on site averages with that based on yearly data. The results of the model are similar (see Figure 4-6 shows an example of a prediction at one site).

**Table 4-3**
**Comparison of Site Average Prediction Equation Coefficients to Those Based on Yearly Site Data (the Coefficients are for $SARFI_{ITIC}$ Predictions)**

| | $k_0$ Intercept | $k_1$ $l$ | $k_2$ $N_g$ | $k_3$ $\frac{n_f \cdot kV^2}{MVA_{xfmr}}$ | $k_4$ Tree term | Dispersion |
|---|---|---|---|---|---|---|
| Site average | 4.26 | 0.480 | 2.20 | 0.247 | 6.33 | 0.389 |
| Yearly data | 3.22 | 0.460 | 2.61 | 0.269 | 3.34 | 0.527 |

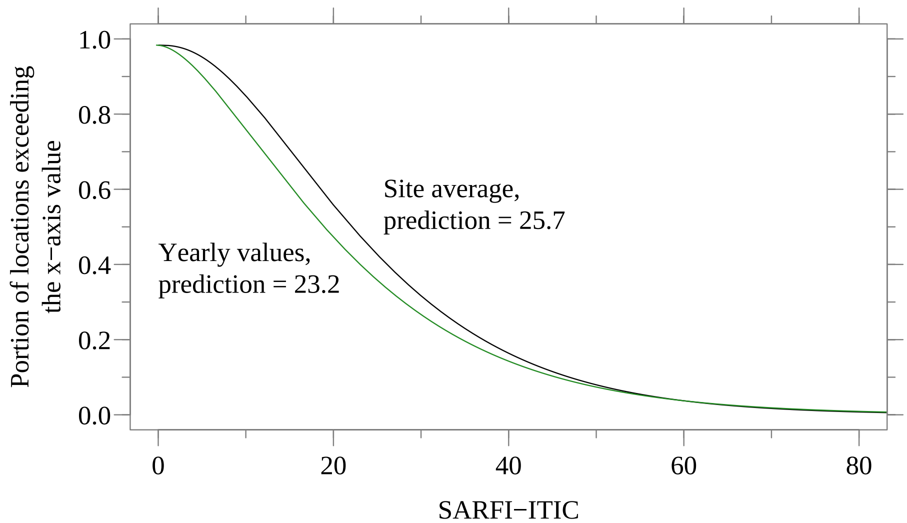



**Figure 4-6**
**Example of Yearly Predictions Versus Site Average Predictions**

We might want to use this model instead of the site average model, given that we often want to know variability in a yearly measurement (especially if goals or performance-based goals are set on a one-year time period). We have to be careful because the yearly data are not actually based on a year's worth of measurements—it is less because of monitor unavailability. Table 4-4 compares the durations of monitoring based on yearly data and site averages. So, our "yearly data" are somewhere in between the site averages and the yearly data results.

**Table 4-4**
**Comparison of the Monitoring Durations for the Site Averages and Yearly Site Data**

| | Monitoring duration, years | |
|---|---|---|
| | Average | Median |
| Site average data | 1.7 | 1.8 |
| Yearly data | 0.7 | 0.8 |

Therefore, we need to be careful in interpreting our prediction results. The "site-average" predictions and prediction bands are estimates of short-term voltage-sag rates. The "yearly-data" predictions and prediction bands are estimates of shorter-term voltage sag rates. Fortunately for many applications, the two results are close enough to each other that the prediction bands and predictions do not change enough to warrant major concern.

A more rigorous evaluation of the variability would help add precision. This is difficult to do given the problem of relatively short monitoring periods and periods of unavailability. The best way to tackle the problem would be with a bootstrapping approach. One way to do it would be to bootstrap yearly estimates by picking from a sample of daily SARFI values from the dataset and derive a spread of yearly data from that.

## Other Ways to Estimate Voltage Sags

The procedure documented in this report is not the only way to estimate the frequency of voltage sags at a site. The prediction models described here are useable with a few easily available inputs. Keep in mind that even though the site parameters are important electrical characteristics, the prediction model is not based on an electrical model—it is a derived regression model. As such, it is difficult to use for analyzing the effect of some changes on voltage sags such as faster relaying, more tree trimming, animal guards, and so on. Another common way to estimate voltage sags is using a model of the system and a short-circuit program. We briefly discuss two variations of this.

The calculation of the voltage magnitude at various points on a system during a fault at a given location is easily done with any short-circuit program. We make the fairly accurate assumption that the fault impedance is zero. The engineer or computer program finds the duration of the sag using the time-current characteristics of the protective device that should operate along with the fault current through it.

Based on a short-circuit program, the *fault positions method* repeatedly applies faults at various locations and tallies the voltages at specified locations during the faults. The runs, which may apply thousands of fault locations, result in predictions of the number of voltage sags below a given magnitude at the specified locations. This procedure is well documented in the IEEE Gold Book[2] (see also Conrad et. al.[3]).

The faults are applied along each line in a system. The end results are scaled by the fault rate on the line, which can be based on historical results or typical values for the voltage and construction.

We need considerable detail for the fault-positions analysis, especially a complete system model including proper zero-sequence impedances and transformer connections (these are left out of many transmission system load-flow models).

Another simpler method for voltage sags is the *method of critical distances* (Bollen[4]). The approach is to find the farthest distance, the *critical* distance, to a fault that causes a sag of a given magnitude. Pick a sag voltage of interest, 0.7 per unit for example. Find the critical distance for the chosen voltage. Using a feeder map, add up the circuit lengths within the critical distance. Multiply the total exposed length by the fault rate—this is the number of events expected. This method is not as accurate as the fault positions method, but is much simpler: We can calculate the results by hand, and the process of doing the calculations provides insight on the portions of distribution and transmission system that can cause sags to the given customer. We can also target this *area of vulnerability* for inspection or additional maintenance or apply faster protection schemes covering those circuits (to clear faults and sags more quickly).

[2] IEEE Std. 493-1997, *IEEE Recommended Practice for the Design of Reliable Industrial and Commercial Power Systems (Gold Book)*.

[3] Conrad, L., Kevin, L., and Cliff, G., "Predicting and Preventing Problems Associated with Remote Fault-Clearing Voltage Dips," *IEEE Transactions on Industry Applications*, vol. 27, pp. 167–72, 1991.

[4] Bollen, M. H. J., *Understanding Power Quality Problems: Voltage Sags and Interruptions*, IEEE Press, New York, 2000.

Whether using the critical distance or the fault positions method, the key input—and the most unknown parameter—is the fault rate. Rates of voltage sags are directly proportional to rates of faults. Good results from the circuit modeling approaches require good estimates of fault rates. Estimates of the uncertainty and variability associated with a prediction must come from knowledge of the variability of fault rates.

Both the circuit modeling approach and the statistical prediction approach described in this report have their uses. The statistical approach has the advantage that it is fast and easy (at least the plain prediction equation without the bootstraps). We can also crudely use it to evaluate some simple changes in circuits such as increasing or decreasing the number of feeders from the bus. The statistical approach is also based on real data and easily provides a good idea of the prediction band within a given confidence level, and with some extensions, the statistical approach can incorporate measurement data. The circuit modeling approach has the advantage that we can use it to estimate the effect of more precise circuit changes: moving circuit sections, adjusting voltage regulation, increasing source stiffness, and so on. The circuit modeling approach is more work and requires much more data input. For some applications, both methods may be appropriate—the statistical prediction gives a quick answer and a good idea of the variability; the circuit modeling approach can be used to evaluate a wider range of solution options.

# 5
# SUMMARY

## Predictions

The main result of this analysis is the SARFI predictions, which fit the following equation with the coefficients given in Table 5-1:

$$SARFI_X = k_0 + k_1 l + k_2 N_g + k_3 \frac{n_f \cdot kV^2}{MVA_{xfmr}} + k_4 \text{ if moderate to heavy tree coverage} \qquad \textbf{Eq. 5-1}$$

where,

$SARFI_x$ = predicted annual number of events that sag below the given SARFI criteria

$l$ = total exposure (including three-phase and single-phase portions) on the circuit in miles

$N_g$ = lightning ground flash density in flashes/km$^2$/year (see Figure 2-3)

$kV$ = base line-to-line voltage in kV

$n_f$ = total number of feeders off of the substation bus

$MVA_{xfmr}$ = station transformer base rating (open-air rating) in MVA

**Table 5-1**
**Coefficients for Prediction Equations for Various SARFI Indices**

| | $k_0$ Intercept | $k_1$ $l$ | $k_2$ $N_g$ | $k_3$ $\frac{n_f \cdot kV^2}{MVA_{xfmr}}$ | $k_4$ Tree term | Dispersion |
|---|---|---|---|---|---|---|
| $SARFI_{80}$ | 11.99 | 0.426 | 2.71 | 0.322 | 3.55 | 0.399 |
| $SARFI_{ITIC}$ | 4.26 | 0.480 | 2.20 | 0.247 | 6.33 | 0.389 |
| $SARFI_{70}$ | 6.23 | 0.353 | 2.24 | 0.195 | 4.37 | 0.435 |
| $SARFI_{SEMI}$ | 3.50 | 0.356 | 1.56 | 0.135 | 6.18 | 0.436 |
| $SARFI_{50}$ | 4.33 | 0.358 | 1.64 | 0.000 | 3.89 | 0.552 |

Keep in mind that these are not precise estimates, but they do help more accurately define the probability range at a given site. The dispersion term of the model can be used to find prediction limits from the Gamma distribution.

## Future Work

Additional work in the area of prediction of voltage sags is possible. Some areas ripe for exploration are:

*Variability* – One of the big concerns when deriving a prediction at a location is estimating the variability. The prediction models derived in this report have a variability term. This term includes time variability, site-to-site variability, and uncertainty in the modeling. It would help if we could derive a better estimate of the time variability from the data and from the predictions. That would help when quantifying the ranges of possibilities when considering things like performance-based contracts. One way to advance this work is to use bootstrapping (resampling) of the combined DPQ dataset to quantify the time variability of voltage sags.

*Combining reliability indices* – Utilities widely record reliability indices such as SAIDI, but most have few power quality monitors. It may be possible to refine voltage-sag predictions using reliability indices. In a statistical prediction model, the reliability indices would become another site characteristic. This could improve power quality predictions and may also reveal interesting practical ties between power quality and reliability.

*Combining circuit modeling approaches with statistical regression* – Circuit modeling approaches such as the fault-positions method have the advantage that we can use them to evaluate many system reconfigurations. One way to improve the quality of predictions from circuit modeling methods is to use statistical predictions for fault rates—the primary input into the circuit modeling method. A portion of the DPQ Phase II data has some very good fault-rate data as well as voltage-sag data. If we can perform the same type of regression based on various site characteristics, we could predict fault rates in the area where we want to predict voltage sags. Then we could use that prediction as the input to the fault-positions method.

*User-accessible implementation* – The basic prediction models derived in this study can be easily implemented in a spreadsheet, so they are accessible to a wide range of users. But some of the more advanced prediction options are difficult or impossible to implement in a spreadsheet. If a user has unknown data for some of the site parameters or wants to incorporate measurements, these are beyond simple spreadsheet implementation. A more user-accessible implementation of the more advanced predictions is possible. The best option for this is probably a Web-based user interface to a calculation engine running R or S-PLUS, where the codes for the advanced modeling are already implemented.

# A
# PREDICTION APPROACH BASED ON DPQ I DATA

The following section is adapted from these sources:

1. T. A. Short, Electric Power Distribution Handbook, CRC Press, 2004.
2. Analysis of Extremely Reliable Power Delivery Systems: A Proposal for Development and Application of Security, Quality, Reliability, and Availability (SQRA) Modeling for Optimizing Power System Configurations for the Digital Economy, EPRI, Palo Alto, CA: 2002. 1007281.
3. T. A. Short, A. Mansoor, W. Sunderman, and A. Sundaram, "Site Variation and Prediction of Power Quality," IEEE Transactions on Power Delivery, vol. 18, no. 4, pp. 1369–75, October 2003.

## Factors That Influence Sag and Momentary Rates

Power system faults cause voltage sags and momentary interruptions. The frequency of faults depends on many factors, including weather, maintenance, and age of equipment. The protection schemes and location of circuit interrupters determine whether a fault causes a voltage sag or an interruption, and the protection system determines the event duration.

### *Location*

Three monitors were used on each circuit in the DPQ study. One was always at the substation, and two were on the feeder, named "feeder middle" and "feeder end." The feeder sites were randomly picked on the circuits, so the naming is somewhat misleading; "feeder end" does not mean the most distant point from the substation (it just means the most distant of the two monitors randomly placed on the circuit). Because one third of the monitors are at the substation, the set is biased to "near-substation" customers because most customers are not located near the substation. Although there is some difference between measurement locations, it turns out that it is not drastic. There is surprisingly little difference between the distributions of monitoring locations (see Figure A-1 for $SARFI_{ITIC}$).

Figure A-2 shows a more specific comparison of the substation's performance plotted against its two feeder sites. As expected, most feeder sites have more sags than their substation site, especially rural sites. A significant number of feeder sites were better than the substation. Measurement anomalies could produce this (the substation recorder is down for part of a bad storm season), or it could be real (downstream regulation devices keep the nominal voltage higher or the connected load "pushes" back on the source impedance during bus faults). For most of our analysis, we excluded substation sites, thinking that the feeder sites better represent a random feeder location where customers are fed.

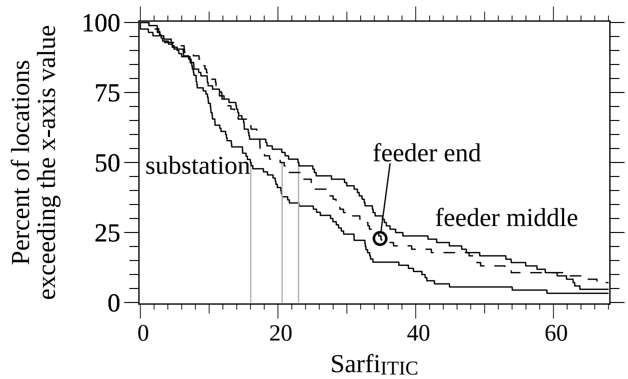



**Figure A-1**
**Comparison of Feeder Sites and Substation Sites in the DPQ Data for SARFI$_{ITIC}$**

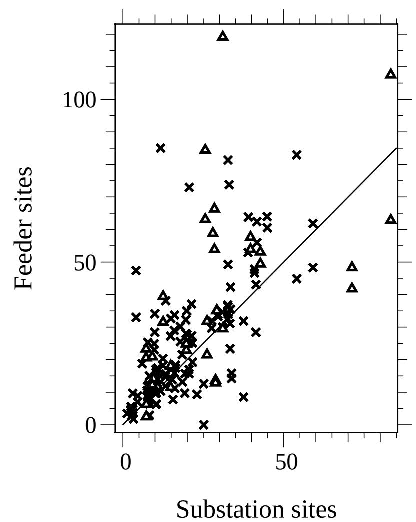



**Figure A-2**
**SARFI$_{ITIC}$ at Substation Sites Plotted Against SARFI$_{ITIC}$ at That Substation's Feeder Sites (Triangles Indicate Rural Sites)**

### *Load Density*

Rural sites have more voltage sags and momentary interruptions (see Table A-1 and Figure A-3). This is not surprising given the extra lengths of line needed to serve load in low-density areas. Interruptions showed the most dramatic difference.

**Table A-1**
**Statistics for Momentary Interruptions Longer Than 0.4 Seconds**

| | P(75%) | Median P(50%) | P(25%) |
|---|---|---|---|
| Rural | 2.37 | 8.56 | 18.31 |
| Suburban | 0.23 | 2.39 | 6.71 |
| Urban | 0.00 | 1.37 | 2.82 |

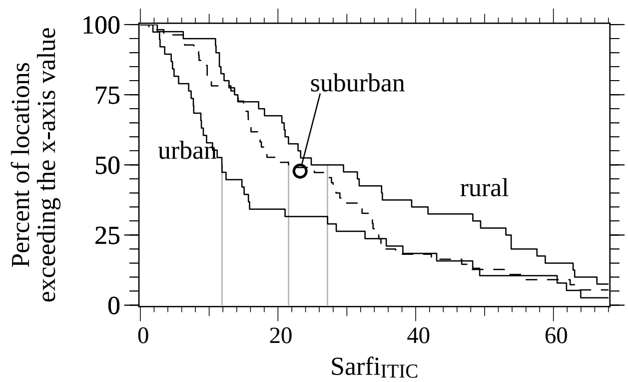



**Figure A-3**
**Comparison of Urban, Suburban, and Rural Sites for $SARFI_{ITIC}$ (Feeder Sites Only)**

Why don't urban sites have even more profoundly lower voltage-sag rates than suburban and rural sites? After all, urban sites are shorter and mostly underground (fewer faults per mile). The main answer is that urban sites have many more feeders off a bus. In addition, even though urban circuits are shorter, most of the exposure is close to the substation. So, while many of the faults on rural and suburban circuits are too far away to pull down the substation voltage, almost every fault on an urban circuit causes a significant voltage sag for all customers off that substation bus.

### *Voltage Class*

Figure A-4 shows that 5-kV systems have much lower numbers of voltage sags and interruptions. Lower-voltage systems have less feeder exposure and higher line impedance (relative to the station transformer). Fault rates are often lower on 5-kV systems. Somewhat surprisingly, the 25- and 35-kV systems were not worse than the 15-kV systems.

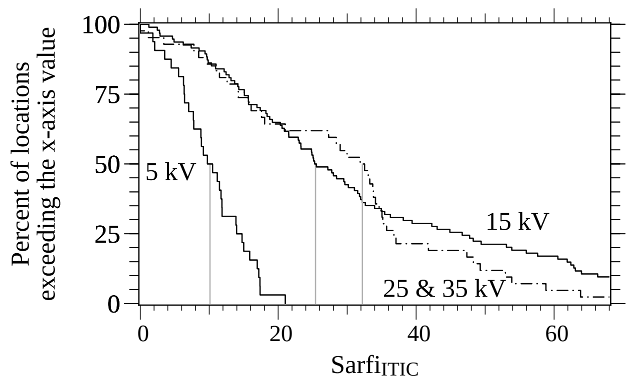



**Figure A-4**
**Comparison of Feeder Sites by Voltage Class in the DPQ Data**

## *Comparison and Ranking of Factors*

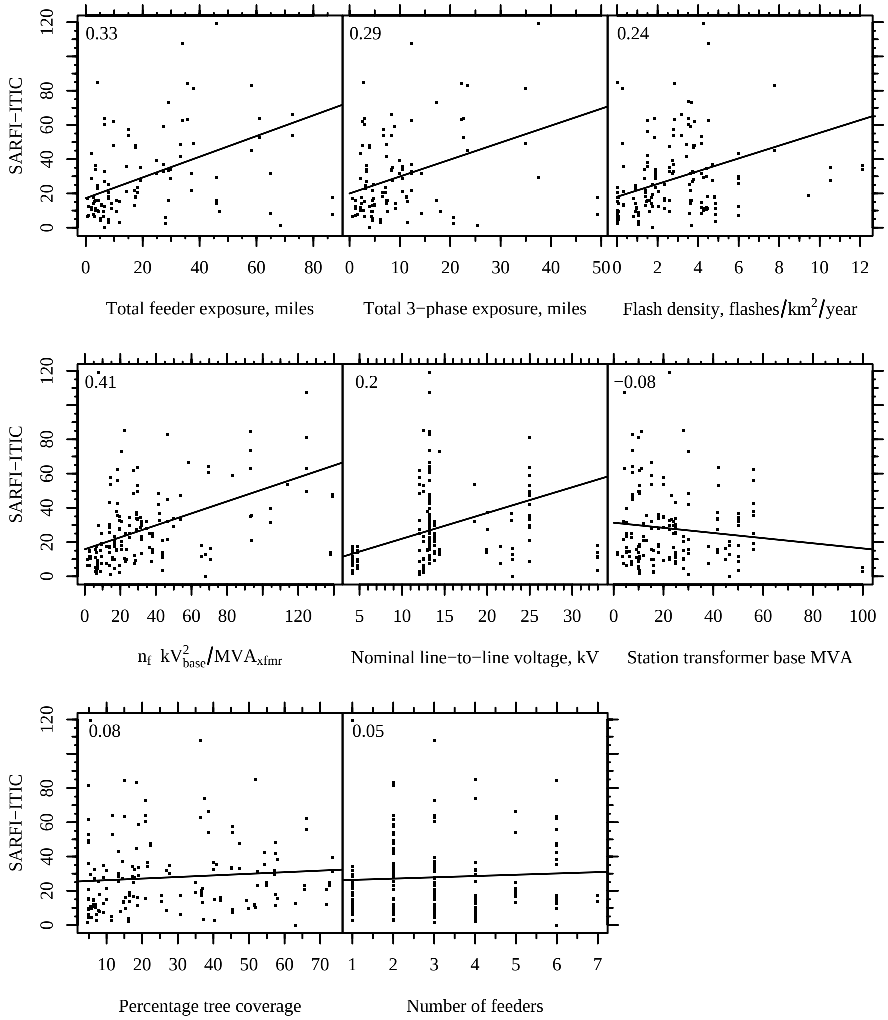
We analyzed data available on the DPQ site characteristics to determine what parameters most affected power quality events. Figure A-5 shows the variations of $SARFI_{ITIC}$ with site characteristics.

The three most significant predictors of excursions below the lower ITI curve are:

1. *Circuit exposure* – The total exposure on the circuit, including three-phase and single-phase portions, is a good predictor of voltage sags. Any fault on the circuit sags the voltage.

2. *Lightning* – Lightning causes many faults on distribution systems, and lightning strongly correlates with voltage sags (based on the ten-year average [1988–98] from the U.S. National Lightning Detection Network). In addition, lightning predicts weather patterns—areas with high lightning tend to have more storms and more wind and tree-related faults.

3. *Transformer impedance and number of feeders* – The $n_f \cdot kV^2/MVA_{xfrmr}$ term in Figure A-5contains the number of feeders off the transformer bus along with an estimate of the transformer impedance. The transformer impedance is $Z_{\%} kV^2/MVA$; but because the per-unit impedance of station transformers is roughly constant (7 to 10%), we use $kV^2/MVA$.




The correlation coefficients (*r*) are given in the upper-left corner of each plot.

**Figure A-5**
**Variations in the Number of Excursions Below the Lower ITI Curve (Mainly Voltage Sags) Versus Various Site Parameters**

This last term requires a bit more explanation. The number of bus sags is directly proportional to $n_f$, the number of feeders off the bus, and to $Z_S$, the source impedance (a lower station transformer impedance, a bigger transformer or lower percent impedance, improves voltage sags at the station bus). We approximate these two terms as $n_f kV^2/MVA_{xfrmr}$.

Other variables have much less impact on the number of voltage sags than the three main parameters given.

## Prediction of Quality Indicators Based on Site Characteristics

We derive a formula for predicting the number of events for a quality indicator based on a few of the characteristics of the site. If no measurement or historical data is available, this is useful in estimating the utility-side quality.

Regression techniques are commonly used to find a model prediction formula. A generalized linear model is a least-squares fit to an equation of the following form:

$$y = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n + \varepsilon \qquad \textbf{Eq. A-1}$$

The *x*'s are site characteristics (such as base voltage or lightning flash density), and the *a*'s are coefficients fitted to the model. The generalized linear model is somewhat different from a standard linear model; we used a generalization where the distribution of the error ε is assumed to be a Gamma distribution rather than a Normal distribution in a strictly linear model. A Gamma distribution skews to the right, like the lognormal distribution.

A model for estimating $\text{SARFI}_{\text{ITIC}}$ is:

$$N_{ITIC} = 4.74 + 0.293l + 2.47N_g + 0.192\frac{n_f \cdot kV^2}{MVA_{xfmr}} + 8.2 \text{ if moderate to heavy tree coverage} \qquad \textbf{Eq. A-2}$$

where,

$N_{ITIC}$ = predicted annual number of events which fall under the lower ITI curve

$l$ = total exposure (including three-phase and single-phase portions) on the circuit in miles (multiply kilometers by 1.609)

$N_g$ = lightning ground flash density in flashes/km$^2$/year

$kV$ = base line-to-line voltage in kV

$n_f$ = total number of feeders off the substation bus

$MVA_{xfmr}$ = station transformer base rating (open-air rating) in MVA

If any of the circuit characteristics are unknown, we could use the following medians from the DPQ data:

$l$ = 14.5 miles (23.4 km)

$N_g$ = 2.57 flashes/km$^2$/year

$$\frac{n_f \cdot kV^2}{MVA_{xfmr}} = 25$$

All three variable terms in the linear regression are significant to at least 99% (there is less than a 1% chance that the terms of the model do not influence the prediction). The tree coverage term is less certain—there is a 9% chance that the term is not significant. We based the tree coverage term on the University of Maryland's Global Land Cover Facility data from the Advanced Very High Resolution Radiometer (AVHRR). Half of the DPQ sites had more than 19% of the land area covered by trees, which we defined as "moderate to heavy tree cover."

How good is the model? It is decent given all the factors that affect sags and momentary interruptions and inherent variability. Given the variability of power quality events, it is surprising that the model is this good. Thirty-four percent of the values are within 25% of the prediction, and 60% of the values are within 50% of the prediction. See Figure A-6 for the prediction scatter.

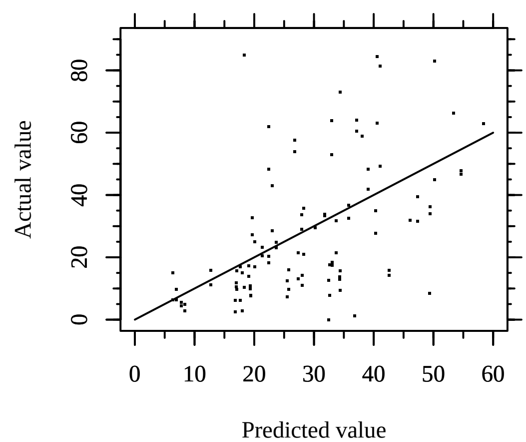



**Figure A-6**
**Actual Values Versus Predicted Values for the Model Predicting the Annual Average Number of Events Below the Lower ITI Curve**

For an example 12.47-kV case with three feeders, a 25-MVA transformer, a flash density of 4 flashes/km$^2$/year, moderate tree coverage, and a total exposure of 32 km, the model predicts 29.8 events per year. For this case, the data show a prediction interval with a 50% confidence level of between 15.6 and 34.2 events per year (the 90% confidence prediction interval is between zero and 68.3). The data are dispersed enough that the model is not good enough to use for precision estimates (such as in a contract for premium power).

The site characteristics most affecting sags but not included in this model (because no information was available) are 1) subtransmission exposure and characteristics and 2) percentage of the circuit that was underground.

A reasonable model for predicting momentary interruptions is:

$$N_{10} = \begin{pmatrix} 5.52 \text{ if Rural} \\ 0.29 \text{ if Suburban} \\ -1.61 \text{ if Urban} \end{pmatrix} + 0.116 l_3 + 0.27 N_g + 1.24 \frac{n_f \cdot kV}{MVA_{xfmr}} \qquad \textbf{Eq. A-3}$$

where,

> $N_{10}$ = the predicted annual number of events with voltage less than 10% of nominal for more than 0.4 seconds
>
> $l_3$ = the three-phase circuit exposure, miles

The parameters differ somewhat from $SARFI_{ITIC}$ predictors. Two of the strongest indicators of momentary interruptions are load density and three-phase circuit exposure. Other significant parameters are the lightning activity and a term with voltage, number of feeders, and transformer MVA. The model is not as good as the $SARFI_{ITIC}$ model, but all parameters have more than a 95% probability of affecting the result. The site characteristic most affecting momentaries that is not included in the model for lack of information is whether fuse saving is used.

*Program:*
Power Quality Solutions

**About EPRI**

EPRI creates science and technology solutions for the global energy and energy services industry. U.S. electric utilities established the Electric Power Research Institute in 1973 as a nonprofit research consortium for the benefit of utility members, their customers, and society. Now known simply as EPRI, the company provides a wide range of innovative products and services to more than 1000 energy-related organizations in 40 countries. EPRI's multidisciplinary team of scientists and engineers draws on a worldwide network of technical and business expertise to help solve today's toughest energy and environmental problems.

EPRI. Electrify the World



*Printed on recycled paper in the United States of America*

1002199

EPRI • 3412 Hillview Avenue, Palo Alto, California 94304 • PO Box 10412, Palo Alto, California 94303 • USA
800.313.3774 • 650.855.2121 • askepri@epri.com • www.epri.com