NASPI-2017-TR-002
PNNL-26313

PMU Data Quality:

A Framework for the Attributes of PMU Data Quality and a Methodology for Examining Data Quality Impacts to Synchrophasor Applications

NASPI PMU Applications Requirements Task Force

March 2017
Version 1.0

Disclaimer

Acknowledgments

This paper was written by Laurie E. Miller (Pacific Northwest National Laboratory [PNNL]), Alison Silverstein (Project Manager, NASPI), Dhananjay Anand (National Institute of Standards and Technology [NIST]), Allen Goldstein (NIST), Yuri Makarov (PNNL), Frank Tuffner (PNNL) and Kevin Jones (Dominion Virginia Power). This effort was supervised and managed by Jeffery Dagle, P.E., and Eric S. Andersen, PMP, both with PNNL.

Funding was provided by the Office of Electricity Delivery and Energy Reliability, United States Department of Energy.

Reviewers and contributors included:

- Ken Martin (Electric Power Group)
- Pavel Etingov (PNNL)
- Harold Kirkham (PNNL)
- Anthony Faris (Bonneville Power Administration)
- Simon Mo (Electric Power Group)
- William Wagner (Bonneville Power Administration)
- Mladen Kezunovich (Texas A&M University)
- Terry Jones (Oak Ridge National Laboratory)
- Ryan Nice (PJM Interconnect LLC)

Executive Summary

With the proliferation of phasor measurement unit (PMU) devices across North America's bulk power system, there is growing use of PMU data in applications that enhance grid operations, analytics, and strategic planning. Some of these applications are critical and require accurate, reliable data, delivered on-time. However, PMU data are often delivered to applications with a variety of data quality issues. These problems are due to the combinations and interactions of the many components of the end-to-end synchrophasor measurement and data delivery process. It is clear that the overall quality of the data used in these applications significantly impacts the accuracy and trustworthiness of the output. However, to date, limited effort has been made to characterize and understand the nature of these impacts. To characterize and understand these impacts, each application must be evaluated using a consistent methodology with common terminology and practices.

The phrase 'data quality' has become a colloquialism that encompasses a broad spectrum of technical meaning and implication. In general, data quality encompasses the accuracy, availability, timeliness, and "fitness for use" of data. However, using a common, specific, clear, and concise language to describe data quality could enable better solutions to PMU data quality problems. This report provides the structured set of terminology needed to describe the many relevant quality-based attributes of PMU data.

In general, the framework laid out in this paper categorizes the attributes of PMU data quality in terms of the scope of the data units being described:

- **Single data points** – attributes of individual data points mostly involve accuracy and metadata.
- **Collections of data points** – attributes of a dataset include data coverage (e.g., time and topology) and consistency of the dataset (e.g., consistent metrology, headers, and adherence to standards).
- **Data streams** – Attributes of a dataset in motion consider the process path and availability.

This report offers detail on these attributes, provides methods for illustrating and representing most data attributes, and identifies where different data problems may arise. Table ES-1 shows the high-level breakdown of the attributes and problems described.

Table ES-1: Attribute Categories

Attributes of single data points
Measurement specifiers
Measurement accuracy
Attribute accuracy
Data lineage (i.e., metadata)
Attributes of datasets
Data lineage (i.e., metadata)
Logical consistency
Data completeness
Characteristics of the data process path
Attributes of an incoming PMU data stream
Characteristics of the data process path
Data stream availability

In addition, this report proposes a methodology for determining the impact of data quality on application performance. This methodology entails the following:

- **Benchmarking** – testing an application many times using many datasets with known errors relative to a clean dataset(s), effectively creating a multi-dimensional PMU data error analysis. Benchmarking allows determination of how different errors affect an application's performance.
- **Standardization** – using the data error analysis results to develop an application-specific performance envelope that shows the characteristics and magnitude of data errors which can be tolerated and yet deliver sufficiently accurate output (based on users' or developers' acceptability requirements), as well as which data errors can render the application's output inaccurate or untrustworthy.

This comprehensive methodology will allow the synchrophasor community to distinguish whether improvement is needed from PMUs, networks, or aggregators. It could also be used to develop a real-time application evaluator that warns when the incoming data stream renders the application untrustworthy. Eventually, this methodology may help application engineers and developers create applications that are less vulnerable to various data quality problems.

Table of Contents

1.0 Introduction

Many transmission owners, generator owners, and balancing authorities have installed phasor measurement units (PMUs) and adopted synchrophasor applications to advance grid analytics, planning, and operations. However, PMU data[1] delivered to synchrophasor applications may vary widely in quality and the condition of PMU data may diminish as it flows through the data process path[2] from point of measurement through communications networks to point of use. To date, there is not a clear and systematic understanding of how problems that develop in the data process path may affect the performance of various synchrophasor applications,[3] and thus the ability of those applications to serve users' needs and goals effectively.

This study looks at data issues from two perspectives:

- What happens to the data as it flows up from measurement at the PMU to the application, and how can users characterize data? Problems in data accuracy, quality, and availability may arise at many points along the path from the PMU to an application. This paper offers (1) a set of terms and definitions to describe and characterize various data problems and (2) a methodology to determine the acceptable performance envelope of an application in terms of PMU errors and data weaknesses. More specifically, this paper aims to facilitate consideration and agreement on how to characterize, define, and measure PMU data flaws, including the accuracy, quality, and availability of PMU data individually and collectively, both in archived or stored form and in data streams.
- How do users determine the requirements for data quality and "fitness for use" for a specific application? Many synchrophasor applications are limited by the quality of the data they receive.[4] To produce accurate, credible results, real-time applications (e.g., voltage stability analyses) are likely to have much more stringent requirements for data quality, completeness, and timeliness than slower applications (e.g., frequency monitoring) or off-line applications (e.g., model validation). Thus, this document offers a starting point for how to understand and assess the impact of these data problems on the performance of synchrophasor applications—in other words, how to determine an application's requirements for input data and assess a data stream's "fitness for use" for that application.

First, NASPI's PMU Application Requirements Task Force (PARTF) conducted extensive research into the definition and categorization of data quality and availability within the synchrophasor field and in other fields that have substantial areas of technical overlap with the field of PMU data generation, transport, storage, and its many uses. Technical areas that had definition sets that matched these criteria included information technology, geographic information systems (GIS), and telecommunications. Standards in the synchrophasor domain, as well as communications and measurements, were examined

1 The term "PMU data" is used here to mean data collected by phasor measurement units, including measurements of frequency, power, voltage, and time-synchronized phasors (synchrophasors).

2 The term, "data process path," is used here to indicate the entire path through interconnected devices and communications networks through which data flows from the point of measurement through communications networks to point of use.

3 The Oak Ridge National Laboratory report, "Impact of Measurement Error on Synchrophasor Applications," (July 2015), addresses the impact of phase angle measurement errors (as determined relative to IEEE Standard C37.118.1-2011 maximum error thresholds) on specific events analyzed with event location, event simulation, oscillation detection, islanding detection, and dynamic line rating applications. The report concludes that the sensitivity of these applications to the maximum allowed phase angle measurement error varies by application and case, with dynamic line rating the most sensitive application. This study did not look at the impact of data problems introduced by system elements beyond the PMU measurement.

4 The phrase, "garbage in, garbage out," is widely used for good reason.

for relevant terminology and definitions. The definitions presented in this document are tailored to the PMU data applications field, but historical context and suggested literature are included in Appendix B.

Because different synchrophasor applications have differing vulnerabilities to PMU data accuracy, quality, and availability, a standard methodology would be valuable for identifying the performance envelope within which an individual application is producing trustworthy results. This methodology could be extended to characterize the diminishing quality of the application's results as the flaws increase within the PMU data being used for analysis. The methodology should have the ability to identify data error impacts, leveraging and moving beyond the traditional standards imposed for a particular device or data method.

Ultimately, the results of this inquiry could be used to develop standard tools for data evaluation and monitoring. These concepts might then be used to formalize a common assessment scheme for data quality, to use that scheme to characterize application data requirements, and to design applications that are less vulnerable to data weaknesses (or at minimum, to better understand the vulnerabilities of an application to various data problems).

1.1 Goals for PMU Data

The Phasor Applications Requirements Task Force recommends three goals with respect to PMU data use:

1. **Data Monitoring** – PMU data should be accurately monitored, and the networks, data concentrators, and archives that touch the data along its path from measurement to use should not significantly degrade the quality, completeness, or timeliness of individual and collected PMU data.

2. **Data Management Tools** – Because some data problems seem inescapable, effective processes and tools should be implemented to detect and deal with bad, flawed, or missing data (e.g., by identifying data redundancy, detecting bad data, cleansing, extrapolation, substitution, and using criteria to identify when a subset of data is not good enough to use).

3. **Application and User Expectations** – Applications that use PMU data should be designed so that they are not highly vulnerable to moderate levels of PMU data problems. The nature and impact of application data vulnerabilities should be understood and factored into users' expectations and trust. In addition, synchrophasor applications should monitor and warn the user when data problems actively compromise the quality of the analytical result.

1.2 Organization of the Report

This report is organized as follows:

- **Section 2: Data Attribute Terminology and Definitions** – this section starts by describing the points along the path from measurement to use where PMU data could be degraded. It lays out proposed terms and definitions for data attributes, offering ways to characterize and evaluate data, regardless of its intended use. In addition, this section describes a method to explain and illustrate many of the different types of data attributes. A detailed table in Appendix A lists all of these terms, along with definitions and the metrics for each, and links those terms back to the data process path.
- **Section 3: Looking at Data from the Application Requirements Viewpoint** – this section lays out application-based data requirements to better address a dataset or stream's "fitness for use" for a specific application.

- **Section 4: The Impacts of Data Issues on Specific Applications** – this section proposes a method for determining how different types of data flaws affect the performance of a specific application.

2.0 Data Attribute Terminology and Definitions

The synchrophasor community has discussed PMU data accuracy, quality, and availability for several years without specifically defining these terms. These terms, and others, are often used interchangeably when describing vastly different components of the larger data quality problem. Broadly speaking, this report recommends that two of these aforementioned terms be adopted along with specific, precise meanings. The term *data quality* should be used as an umbrella term that encompasses the accuracy, content, "fitness for use", and availability of data, and *data stream availability* should be used as another umbrella term that pertains to whether relevant data reach the application at all, even if not in time for constructive use.

To provide even more structure to the way the data quality problems are described, it is useful to break down the spectrum of data quality according to the scope of the data units being described. *Data accuracy* issues pertain to individual data points and often arise at the point of measurement, while additional data attributes are defined to describe datasets. In contrast, most *data stream availability* issues pertain to data in motion that is arriving as it is processed. Therefore, this report organizes data attributes as shown in Figure 2-1, recognizing that datasets are collections of data points and data streams are datasets in motion.

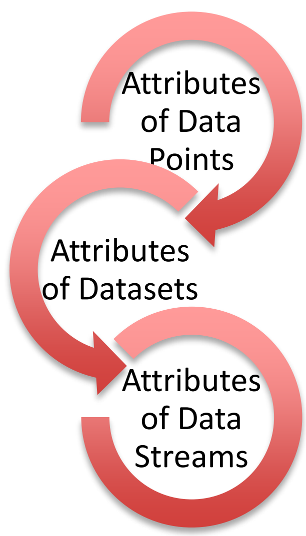


Figure 2-1: Defining Data Quality Attributes in Terms of Their Application to Data Points, Datasets, and Data Streams

These sets of attributes are defined as follows:

- <u>Attributes of single data points</u> enable the characterization of the fitness-for-use of individual measurements for a given purpose. They describe characteristics that should be evaluated for individual data points to determine the data's fitness-for-use for a given purpose. Single data points are a basic representation of the information synchrophasor devices provide. Many attributes at this level are focused on the measurement aspect of the data as well as data descriptors or "metadata" (e.g., geospatial location of the point of measurement).

- Attributes of datasets enable the characterization of the fitness-for-use for collections of single data points for a given purpose. They describe additional identifying information of the dataset as a whole. The simplest form of a dataset is a time-series collection of single data points representing how the result of the measurement changes with time. Datasets can be defined based on the architecture of the measurement delivery system (e.g., the datasets assembled by phasor data concentrators [PDCs] and archives) or by the application (i.e., the set of critical measurements required for basic functionality of the application).
- Data stream attributes pertain to data that is moving through the communications and process network in real time, making attributes such as timeliness and dependability relevant.

In terms of attributes, the boundaries between single data points, datasets, and data streams are not absolute. For example, individual data points have attributes that specify what each datum represents. When individual data points are collected into a dataset, the data point attributes will still apply, but additional attributes are needed to specify what the dataset represents and to characterize its fitness-for-use. Furthermore, the attributes pertaining to individual data points and datasets still apply when those are aggregated into a data stream, with further attributes to characterize that data stream. This attribute overlap is illustrated in Figure 2-1.

2.1 The Data Process Path

PMU data originate at the point of measurement and pass along a process and communications path to the point of use. This is defined as the "data process path." Problems that develop along the data process path may affect the performance of various synchrophasor applications,[5] and thus the ability of those applications to serve users' needs and goals effectively. Conceptually, PMU data problems can arise in four places along the data process path from point of measurement through communications networks to point of use:

- At or within the PMU, from sources that may include biases introduced by the potential transformer (PT) or current transformers (CT) to which the PMU is connected, the class of PT or CT utilized (protection vs. measurement), differences or errors within the PMU and its filters and algorithms, or erroneous or missing timestamps associated with the measured grid conditions. These factors affect how the individual data point is created and whether it is accurate.
- At any intermediate data aggregators (e.g., PDCs, historians, databases, or archives) that receive, process, or store a data point on its trip from the point of measurement to the point of use,[6] as from misalignment, erroneous compression, creation of duplicate data, or loss of data due to late deliveries. The aggregators can create additional data errors (e.g., mislabeling) or data losses.

[5] The Oak Ridge National Laboratory report, "Impact of Measurement Error on Synchrophasor Applications," (July 2015), addresses the impact of phase angle measurement errors (as determined relative to IEEE Standard C37.118.1-2011 maximum error thresholds) on specific events analyzed with event location, event simulation, oscillation detection, islanding detection, and dynamic line rating applications. The report concludes that the sensitivity of these applications to the maximum allowed phase angle measurement error varies by application and case, with dynamic line rating the most sensitive application. This study did not look at the impact of data problems introduced by system elements beyond the PMU measurement.

[6] Most applications (especially on-line, real-time synchrophasor applications) incorporate dedicated aggregation or data concentration services within the tool, which review and may modify the data before the application performs its analytical or control functions. Understanding the interactions of all aggregation methods through which the data has passed can be critical in determining data's fitness-for-use by a given application or algorithm.

- Within the communications network that delivers PMU data from the point of measurement to the point of use, as from loss of communications network nodes, excess latency delaying delivery beyond the acceptable time window of the PDC, or data corruption introducing erroneous values.
- At the point of use by the application, as from data storage issues or an insufficient application training period[7] for an application that requires enough historical data to show the application and its users the appropriate range of application solutions and exceptions.

A discussion of the detailed end-to-end PMU measurement system or data process path is helpful in considering the mechanisms affecting synchrophasor data quality. The data process path can be subdivided as shown in Figure 2-2. (Note that the points numbered in Figure 2-2 are explained below.) The factors that cause a measured value to deviate from the signal being measured on the grid are said to "degrade" the measurement.

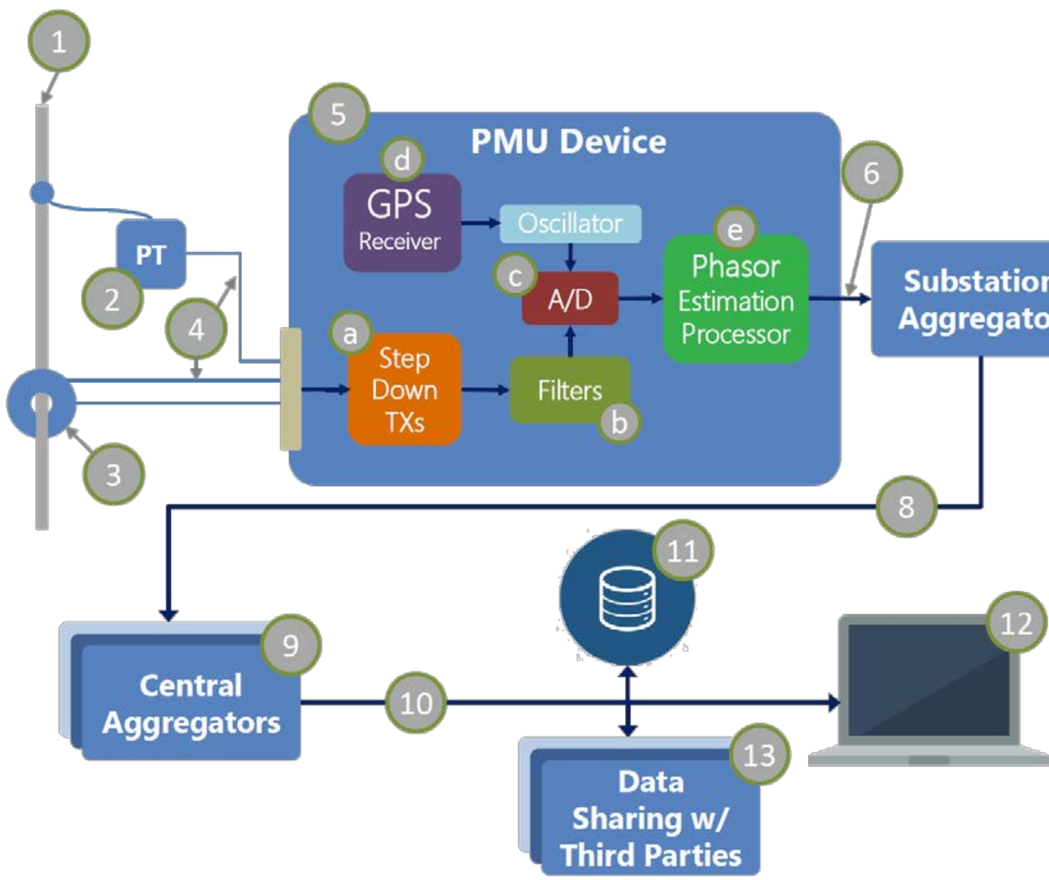


Figure 2-2: The detailed data process path or end-to-end measurement system

[7] The *application training period* and related concepts were discussed under "analytic history" in a previous draft version of this document.

1. **The conductor** – The conductor is the physical transport medium for the electrical energy on the system. The characteristics of the electrical energy being carried contain much more information than is desired. Therefore, the measurement system must be able to differentiate between the desired and undesired information (i.e., noise and other power system phenomena). The potential energy level of the conductor makes direct measurement costly and dangerous. Therefore, these energy levels must be transformed to lower energy levels with instrument transformers (i.e., PTs and CTs).

2. **The potential transformer (PT)** – The PT steps down the conductor potential to a low enough value that it can be read safely and affordably by a PMU. Each PT has a defined step-down ratio for the voltage that defines the relationship between the input and the output. This ratio is an approximation. The PT introduces a complex bias to the measurement that includes a shift in magnitude of the conductor signal and a shift in phase of the conductor signal. With respect to the resultant phasor measurement, these values are called ratio correction factor (RCF) and phase angle correction factor (PACF), respectively.[8] Changes in RCF and PACF can be created by changes in ambient and seasonal conditions as well as changes in the metering equipment condition over its operating lifetime.

3. **The current transformer (CT)** – The CT plays the same role as the PT in the measurement system and in the degradation of measurement quality. A CT measures the ampere flow of the conductor while the PT measures its potential flow.

4. **The instrument transformer secondary signal wire** – The secondary side of each instrument transformer is connected to conductors that carry the signals to the input channels of the PMU device. The impedance of these signal wires may introduce degradation into the measurement quality. If this signal wire is short, this impedance value should be small and, thus, is usually not counted as a measurement adjustment. In addition, other devices connected to the secondary signal wire can change the burden and impact the measurement quality.

5. **The PMU device** – The PMU is the most complicated component of the measurement system and contains the most complex measurement degradation mechanisms. Further, the apparent impedance of the PMU from the instrument transformers introduces error in the signals on the instrument transformer secondary sides; this is called the burden impedance of the instrument transformer. The internal mechanisms of the PMU device affect data quality in the following ways:

    a. **The step-down transformers** – While the conductor signals have already been stepped-down by the instrument transformers, the voltage and current levels fed into the back of the PMU are still too high to directly digitally sample with an analog-to-digital converter. Like the upstream instrument transformers, these smaller step-down transformers will introduce additional complex bias (or more formally represented by another transfer function) into the signals.

    b. **Filters** – The PMU device contains mathematical band-pass and anti-aliasing filters that aid in the selection of the fundamental frequency component that is desired for measurement. These may also introduce non-negligible degradation to the measurement.

    c. **Analog-to-digital (A/D) converter** – The PMU device contains an A/D converter for digitally sampling point-on-wave values that can be used by the signal processor for estimating the

[8] The RCF and PACF values are simplifications of the transformation that takes place within the PT. The relationship between input and output would be better represented with a transfer function representing these shifts across the frequency spectrum of the conductor signal.

phasor values of the signal. The very nature of A/D conversion represents a degradation of measurement information because it truncates and, therefore, quantizes measurement values. In most cases, however, the precision of the A/D converter is designed to introduce negligible error to the measurement relative to the errors of the other system components.

d. **Timing** – The PMU device depends on a high-precision time synchronization mechanism such as that provided by a Global Positioning System (GPS) signal and receiver and a local crystal oscillator. Time synchronization enables measurement timestamping and the accuracy of the phasor value itself. PMU standards specify the required accuracy and precision of time signals and time synchronization. Timestamps within the PMU are affected by errors such as misconfiguration and the loss of satellite signal (if GPS-originated) or network-delivered time. However, timing accuracy is also subject to errors such as radio interference (e.g., from poor GPS receiver placement) and GPS spoofing. A substation's GPS time source may be supplemented with a land- or IP-based time transfer signal.

e. **The phasor estimation algorithm in the microprocessor** – A PMU feeds the samples output by the A/D converter into an algorithm to convert these point-on-wave, timestamped samples into a phasor valued representation of the signal, the estimated "phasor measurement". Although these algorithms are dictated by technical standards, they vary by vendor and sometimes between devices.

6. **The substation local area network (LAN) or substation data bus** – The substation LAN or substation data bus is part of the communication system that delivers PMU data from the point of measurement to the point of use. Communication systems are vulnerable to the loss of communication network nodes, excess latency delaying delivery beyond the acceptable time window of the PDC, or data corruption introducing erroneous values. In-substation communication lines are less likely to introduce measurement degradation if they have been configured properly and remain in a physically secure and environmentally controlled environment.

7. **The substation aggregator** – Many times, a substation PDC or data aggregator may be part of the measurement delivery system. However, some PMU owners connect the PMU device directly to the wide area network (WAN) and thus straight to the central data aggregators. The substation aggregator is subject to measurement degradation mechanisms that include misalignment, erroneous compression, creation of duplicate data, or loss of data due to late deliveries. Substation aggregators can create additional data errors such as mislabeling or data losses.

8. **Wide Area Network (WAN)** – The WAN delivers data from the substation (whether from the PMUs or the substation aggregator) to higher-level PDCs, aggregators, and applications. The WAN can introduce data delays, data drops, and data corruption.

9. **The central data aggregator** – The central data aggregator, often a PDC, introduces similar modes of measurement degradation as the substation aggregator.

10. **The organizational WAN (intranet)** – The PMU data user's intranet could introduce additional latency or data corruption.

11. **Archive/historian aggregators** – Additional aggregators (e.g., historians) can create data loss or reduce data accuracy by modifying arriving PMU data through well-intentioned processes (e.g., data compression, data truncation, or data deletion).

12. **Application/analytic/visualization** – The application itself may modify the data, including pre-processors that may perform some data aggregation functions; the analytic core of the application itself; and transformations of the output for visualization of the results.
13. **Wide area phasor networks** – If a PMU owner passes data to third parties (e.g., independent system operators (ISOs) and regional transmission operators (RTOs) through a hierarchical, rather than parallel, communications network,[9] it may introduce another layer into the data process path through the addition of another set of WAN links and third-party aggregators and applications. This additional layer may introduce further measurement degradation.

2.2 Attributes of Single Data Points

Single data points are one of the most basic representations of the information that synchrophasor devices can provide to applications. At this level of characterization, many of the attributes are focused on the measurement aspect of the data such as the accuracy and precision of the measurement value as well as the metadata (e.g., geospatial location) of the data point. Broadly speaking, the classification of attributes includes three categories: (1) the measurement specifiers, (2), the accuracy of the measurement and the measurement attribute, and the (3) data lineage. These characterizations are summarized in Table 2-1 and detailed in this section along with examples of various attributes and related data errors.

2.2.1 Measurement Specifiers

The first category of attributes for single data points is measurement specifiers (see Table 2-1). Measurement specifiers are the minimum attributes necessary to report a measurement with sufficient clarity. As such, measurement specifiers are attributes that describe the ability of the measured result to properly represent the physical phenomenon being measured. Specifically, this refers to whether the process of measuring some phenomenon on the power system (the signal being measured) and expressing or calculating its value (the measured result) has been specified effectively in terms of standard units to a given precision and within a stated confidence interval. In general, measurement specifiers include several critical attributes:

- **Signal being measured** – A description or definition of the grid condition or quantity that the measurement will represent.
- **Measured result** – this gives the numerical value of the measurement and is further described by the following attributes:
    - **Standard units –** e.g., current measured in amperes.
    - **Precision –** e.g., to three decimal points.
    - **Confidence interval** – e.g., the calculated measurement comes within ±2 percent of the signal being measured.

The values of the measurement specifiers are determined by the components of the measurement system or process that connect the systemic observation of the physical quantity being measured to the transformation of that observation into a unit of information for retention. In the case of PMU measurements, the measurement system primarily runs from the electromagnetic coupling of the instrument transformers with the power system to the output of the discrete Fourier transform performed

[9] In a hierarchical communications network, PMU data flows from the PMU through the substation to the data owner, and then from its aggregators across another network to the regional entity's aggregators and applications. In a parallel communications network, PMU data would flow from the substation onto a single communications network that would deliver data to both the data owner and the regional entity without intermediate processing or delivery steps.

inside the PMU. However, the impact of the measurement system on individual data points can affect downstream components of the measurement system.

<u>A PMU data point is a cluster of measurements</u>

A PMU measurement is clustered, in that the timestamp is an attached attribute measurement that is itself a measurement. In a clustered measurement, each attribute measurement must also characterize the measured result of a signal being measured in terms of standard units to a specified precision within a stated confidence interval.

Table 2-1: Single Data Point Attributes and Metrics

Measurement specifiers		Signal being measured
		Measured result
		Standard units
		Precision
		Confidence interval
Measurement accuracy		Source influence: • Noise in the power system • Out-of-band interfering signals • Harmonics • Instrument channel errors • GPS loss, spoofing, or meaconing[†] Induced error (created by the PMU estimation process): • Synchrophasor total vector error (TVE) • Synchrophasor magnitude error • Synchrophasor phase error • Frequency error • rate of change of frequency (ROCOF) error
Attribute accuracy		Temporal accuracy (timestamp matches time of measurement)
		Geospatial accuracy (coded PMU location matches actual location)
		Topological accuracy (coded PMU topological location matches location in actual power system topology)
Data lineage (i.e., metadata)*	Data source	PMU type
		PMU standard followed
		PMU model, firmware version, configuration settings
		PMU-supplied data headers
		Aggregator-supplied data headers
	Data coverage	PMU geospatial location
		PMU topological location
	Data transformation methods	Transformations applied to the data at the PMU

[†] Meaconing refers to interception of signals and rebroadcast of altered signals in their place.

* Assess regularly – this attribute does not need to accompany every measurement, but should be weighed when determining the suitability of data for an application or study or when troubleshooting problems with application results.

2.2.2 Measurement Accuracy and Attribute Accuracy

The next top-level category of single data point attributes is data accuracy, which can be divided into two subcategories:

- **Measurement accuracy**[10] takes into account the following properties:
  - **Source influence** – created by phenomena within the power system that could cause the measured result to differ from the signal being measured, such as harmonic interference in the power system (see Table 2-2). Source influence deals with the need and the ability to separate the physical quantity of interest from other physical quantities that affect the observation of the desired quantity. In general, this is the *differentiation* of information within the measurement system. In power systems, source influence might include phenomena (e.g., harmonic interference).
  - **Induced error** – the error introduced by the estimation process or methodology within the PMU. In general, this is the *degradation* of information within the measurement system (see Table 2-3).
- **Attribute accuracy** refers to the correctness of other attributes attached to the measurement. In the case of PMU data, this includes the following:
  - **Temporal accuracy** – the agreement between the timestamp assigned to a measurement and the actual temporal coordinates at the instance of measurement. GPS spoofing or a satellite timing error could result in a deviation from the expected *temporal accuracy* of a PMU measurement.
  - **Geospatial accuracy** – the agreement between encoded and actual location coordinates of the PMU on the earth's surface.
  - **PMU topological accuracy** – the agreement between the recorded electrical location of the PMU and its actual electrical location in the topology of the power system.
  - **Measurement topological accuracy** – the agreement between the recorded electrical location of the measurement and its actual electrical location in the topology of the power system. A single PMU may make many measurements at various topological locations (in a substation), so this is not necessarily the same as PMU topological accuracy.

While the timestamp and its temporal accuracy are a part of the PMU data stream, geospatial location and topological location are not usually processed as part of the PMU data stream itself. Topological and geospatial errors are most often due to poor PMU set-up and registration or maintenance. However, geospatial accuracy and topological accuracy can have significant impacts on application output accuracy and trustworthiness if they are not understood and managed correctly. Table 2-2 outlines many of the causes for source influence in PMU measurements, and Table 2-3 itemizes many of the sources of induced errors in PMU measurement.

---

[10] This is referred to as "positional accuracy" in metrology and geographic information systems. It is referred to as "measurement accuracy" in this document to prevent confusion with geolocation measurements ("positions") such as latitude and longitude.

Topological accuracy example

For a transmission line terminating at a switching station, the potential transformer providing the source signal to the PMU may be placed on either side of the line circuit breaker. Therefore, when this line is removed from service, if the PMU voltage reading is on the substation bus it may still show voltage although the line was de-energized. If this is not explicitly understood in the model of what the PMU data is representing, it may indicate an impossible condition and cause errors or misrepresentation in the application.

Table 2-2: Possible PMU Measurement Source Influences

Sources	Measure	Units	Comments
Noise in power system	Wideband signal-to-noise ratio (SNR)	% or dB	
Out-of-band interfering signals	Narrowband SNR	% of nominal magnitude	
Harmonics	Total harmonic distortion (THD)	% of nominal magnitude	
Instrument channel errors	All	Varies	Hardware affecting PMU measurements: • PTs • CTs • Capacitively coupled voltage transformers (CCVTs) • Electro-optical voltage transformers (EOVTs) • Magneto-optical current transformers (MOCTs) • Cables • Differences between measurement-class versus protection-class instrument transformers.
GPS loss, spoofing, or meaconing*			Impact of GPS loss on time drift and measurement error [Liu2015]. This is a type of *attribute accuracy*, specifically *temporal accuracy*.

* Meaconing refers to interception of signals and rebroadcast of altered signals in their place.

Table 2-3: Sources of PMU Measurement-Induced Error

Sources	Measure	Units	Comments
Synchrophasor error	TVE	%	• IEEE Std. C37.118.1-2011, C37.118.1a-2014 • The magnitude of the synchrophasor's root-mean-square (RMS) amplitude/phase error vector
	Magnitude error	%	The difference between the RMS amplitude of the synchrophasor estimate and the RMS amplitude of the PMU's input signal
	Phase error	%, degrees, or radians	The difference between the reported synchrophasor angle and the absolute angle of the PMU's input signal
Frequency Error (FE)	FE	Hz	The difference between the reported frequency and the instantaneous frequency of the PMU's input signal
Rate of Change of Frequency (ROCOF) error (RFE)	RFE	Hz/s	The difference between the reported ROCOF and the ROCOF of the PMU's input signal

In addition to the above sources of induced error, differences between PMU devices can cause inconsistencies between how different PMU devices measure the same grid condition. These can include the following:

- **Physical differences** – these include variations in the manufacturer of the PMU and/or the model of the PMU. There may also be physical differences between 'identical' PMUs due to variations in manufacturing processes; these could include the burden impedance of the internal transformers of the PMU (which step down voltages and currents from the PT/CT secondary levels to milliamps), or variations across different lots of third-party-manufactured components used in manufacturing a given model of PMU.
- **Algorithmic differences** – these include differences in the signal processing components of the software, particularly the phasor estimation algorithm implementation (Huang 2008).
- **Internal configuration differences** – these include differences in the various settings that dictate how the device behaves irrespective of its hardware.
- **Software and firmware differences** – the software and firmware installed on the PMU have a substantial impact on the behavior of the device that cannot always be captured with configuration or algorithmic differences. Every unit within a single owner's fleet of PMUs should ideally use consistent versions of the software and firmware.
- **Differences in the interpretation of PMU standards** – these include differences in the interpretation of the definition of a phasor, the phasor protocols (C37.118), etc. Two devices that comply with a specific PMU standard and definitions may not perform identically across the range of operation of the device.

The impacts of common mode induced error may vary

As noted above, potential error or bias can arise from many causes. If those biases are consistent across all of the PMUs within a single substation (e.g., because they are all connected in the same way to the same PT or CT) or consistent across an entire fleet of PMUs (e.g., because they are all the same model from the same manufacturer running the same set of measurement algorithms), then the biases or error across the set of PMUs will affect every piece of data in the same way and be consistent over time. Because so many synchrophasor applications look for differences and out-of-variance occurrences in the data over time and topology, rather than only at absolute values, a persistent set of consistently biased (but relatively accurate) data should have little effect on most analytical results. However, inconsistent biases – such as those arising from a change in filtering algorithms or a lasting erroneous timestamp (as has occurred following past leap second events) can cause significant differences between PMU measurements within a fleet, and create material analytical errors.

2.2.3 Data Lineage

The remaining categories of single data point attributes from Table 2-1 belong to the larger category of *data lineage*. Data lineage refers to the sourcing of data and the level of trust assigned based on how that data was created. It also encompasses the metadata associated with a specific PMU device. Attributes of data lineage include:

- **Data source**
  - **PMU type** that the data is sourced from (e.g., M- or P-class PMU)
  - **PMU standard followed** – the specific standard (base standard and revision) that the PMU follows
  - **PMU model, firmware revision, and configuration settings**

    - Any labeling or attribute practices of any PMUs or aggregators through which the data passes (i.e., naming conventions), including **PMU-supplied data headers** and **aggregator-supplied data headers**
- **Data coverage**
    - **Geospatial location** of each PMU (e.g., latitude-longitude)
    - **Topological location** of each PMU (e.g., bus-side voltage phasor for line XYZ). Note that there can be many PMUs inside a single substation. Therefore, electrical topology location may vary from device to device inside a single substation.
- **Transformation methods**
    - **Transformations applied to the data at the PMU** (e.g., downsampling, conversion from rectangular to polar, computation of real and reactive power flows from voltage and current phasors).
    - **Transformations applied to the data at an aggregator** and **transformations applied to the data at an archive** (e.g., truncation or conversion from line-to-neutral to line-to-line kV).

Many of the attributes of data lineage may take the form of metadata recorded in a PMU Registry, as shown in the example in Table 2-4. The metadata represented in data lineage do not have to be re-examined with every individual PMU measurement. However, if there is any reason to suspect that incoming or archived single data points demonstrate characteristics of poor data quality, then it is worthwhile to review data lineage elements (e.g., PMU firmware settings, aggregator labeling practices, or data transformations) to be sure that these remain accurate and consistent. In large organizations, field staff may modify or update field PMUs or connected devices without knowledge of the engineers who use the affected PMU data. Furthermore, intermediate components like PDCs or communications interfaces may undergo standard software updates including operating system patches, which could introduce changes in PMU data availability or accuracy. Such scenarios illustrate the need to reassess data lineage attributes should corruption be observed in single data points.

In the context of single data points, the term "transformation methods" only encompasses those transformations that are applied to the data at the PMU. These are typically simple mathematical formulations, such as calculating the phasor representation of complex power by appropriately multiplying a voltage and current phasor together. More complicated transformations can occur, but often require an aggregation or set of data.

Some discussions replace lineage with "standards-based." This implies that if a dataset adheres to a given standard, its lineage is well characterized, well understood, and therefore meets certain minimum requirements.[11] However, a PMU and its data stream can be standards-based, but still introduce data quality problems through poor standard interpretation, implementation, or updating.

---

[11] Standards-based consistency of data may be a few years away in North America, because there are still multiple generations of research- and production-grade PMUs deployed across the grid, built to varying generations of synchrophasor technical standards. Currently, for the latest standards (IEEE 37.118.1 and IEEE 37.118.2), "compliance" to a specific standard does not mean that two measurements of the same condition by two neighbor PMUs will produce the identical measured result due to internal variations in signal processing, filters, and algorithms.

Table 2-4: Excerpt of PJM PMU Registry Metadata Content (PJM2014)

	Entry Type	Value	
Location	Company Name:		Maximum 50 Characters
	ISO/RTO Name:		
	Substation/Plant Name:		Maximum 50 Characters
	Substation/Plant location: Latitude – Longitude::		Maximum 15 Characters
	Time Zone		Time Zone of the Substation. This is different than the data time stamp which for PJM project is in UTC
	Contact Information for Phasor Data:		
	o Name		
	o Phone		
	o Email		
Phasor Measurement Unit Information	PMU IDCODE		As per assigned IDs; see 'ID Code assignment' worksheet for the IDs available for your company's PMU/PDCs
	PMU In Service Date		mm/dd/yyyy
	STN		Maximum 16 Characters; As defined in 'STN and CHNAM' work sheet in this excel file
	PMU Manufacturer		Maximum 50 Characters
	PMU Model (Name and Version)		Maximum 35 Characters
	PMU Firmware Version		Maximum 15 Characters
	PMU Class		'P' or 'M' (2011 std.) or '0' or '1' (2005 std.)
	PMU Reporting Rate		Reporting Rate
	PMU Operational Status		Free form For Example: In Service, On Outage, Not Connected to PDC, In Testing. Maximum 35 Characters
	Time Synchronization Source (for the PMU):		
	o Manufacturer		Maximum 50 Characters
	o Model (Name and/or Version)		Maximum 35 Characters
	Data Sent to - from this PMU (To be repeated for each PDC):		
	o PDC (ID Code)		As per assigned IDs; see 'ID Code assignment' worksheet for the IDs available for your company's PMU/PDCs
	o PDC Type		Substation or Regional or Central (Main)
Voltage Phasor Measurements (To be repeated for each voltage phasor measurement)	CHNAM		Maximum 16 Characters defined as per specifications on sheet 'STN and CHNAM'
	Data Representation (Polar or Rectangular)		PJM prefers Polar representation
	Data Format (Integer or Floating Point)		PJM prefers Integer format
	If Line Side Measurement:		
	o Line Identifier		Free form field. Could be a description or the ID used by the TO to identify the line when the voltage source is on the line side
	Breaker #1 Identifier		Free form field. Could be a description or the ID used by the TO to identify the Breaker
	Breaker #2 Identifier		
	If Bus Side Measurement:		
	o Bus Identifier		Free form field. Could be a description or the ID used by the TO to identify the Bus if the Voltage source is on the Bus side
	Nominal Voltage		For Example: 138 KV or 230 KV or 345 KV or500 KV
	Instrument Transformer:		
	o Device Description		
	o Accuracy Class		
	o Ratio		
	Phasor relation to reference phasor (Due to Δ-Y)		This would be applicable to 25 KV / 23 KV / 12.5KV systems
	Measurement adjustments required (external to PMU):		
	o Scaling Factors		Free Form Field: to identify if there are any scaling adjustments to be done to the measurements. In most cases this should be not applicable
	o Phasor angle adjustment		Free Form Field: to identify if there are any phase angle adjustments to be done to the measurements. In most cases this should be not applicable

2.3 Attributes of Datasets

Individual PMU data points become useful primarily when they are aggregated into datasets. Types of datasets include:

- A time-series set of data points—one of the simplest forms of an aggregate PMU dataset.
- All data generated by a synchrophasor device.
- A data slice—a subset of a dataset including all data across that dataset for a particular time interval, where the time interval may be a single timestamp or a wider interval.
    - A measurement frame generated by a synchrophasor device, containing multiple measurements associated with a single timestamp, represents a *data slice* for that synchrophasor device.
    - A data slice from an aggregator would aggregate data sources from different geospatial or topological locations.
- All data in a storage archive.

In the overall synchrophasor measurement system, single data points are collected into aggregators, starting with aggregation into data frames leaving the PMU, continuing through PDCs or other types of collection points, into storage systems before processing, and into archival storage. Additional aggregation may be done in application front-ends to prepare data for the application's processing algorithms. Table 2-5 provides an overview of the categories of attributes associated with datasets.

2.3.1 Data Lineage

Data lineage is a key attribute of datasets as well as single data points. While there are certain data transformations methods that affect data at the granularity of single data points, aggregating and combining the data may involve further transformations that could affect an entire set of data, thus affecting the results of an application. These transformation methods are still considered part of the overall *data lineage* category but are extended to multiple measured results at the same time. As noted above, data transformation on a dataset may occur at an aggregator or during data archiving.

The *data coverage* sub-attributes of data lineage define metrics on the electrical, spatial, and temporal dimensions the dataset encompasses. This includes the following:

- **Aggregator geospatial coverage** of the geographic area covered by the aggregator
- **Aggregator topological coverage** of the electrical network area covered by the aggregator
- **PMU temporal coverage** of the span of time for which the PMU has data
- **Aggregator temporal coverage** of the span of time for which the aggregator has data

Geospatial coverage indicates the geographic zones or locations from which the aggregator contains data. *Topological coverage* is similar, but is tied to the locations on the actual electrical network the data delivered to the aggregator encompass and indicates exactly what elements of the electrical network can be considered to be covered by the dataset. *Temporal coverage* of both the PMU datasets and aggregator datasets are an indicator of time the device and the equipment it was measuring was in operation. This dictates the time intervals that are available for analysis.

Table 2-5: Dataset Attributes and Metrics

Data lineage (i.e., metadata)*	Data transformation methods	Transformations applied to the data at an aggregator
		Transformations applied to the data during archiving
	Data coverage	Aggregator geospatial coverage
		Aggregator topological coverage
		PMU temporal coverage (time in operation)
		Aggregator temporal coverage (time in operation)
	Data content	Number of PMUs
		Channels provided by the PMU
		Channels provided by the aggregator
		PMU reporting rate
		Aggregator reporting rate
Logical consistency*	Metrology persistence and consistency	PMU metrology persistence
		Aggregator metrology persistence
		Aggregator metrology consistency
	Header persistence and consistency	PMU header persistence
		Aggregator header persistence
		Aggregator header consistency
	Data frame persistence and consistency	PMU data frame persistence
		Aggregator data frame persistence
		Aggregator data frame consistency
	Data frame order consistency	Aggregator data frame order consistency
	Standards compliance persistence and consistency	PMU standards compliance persistence
		Aggregator standards compliance persistence
		Aggregator standards compliance consistency
	Reporting rate persistence and consistency	PMU reporting rate persistence
		Aggregator reporting rate persistence
		Aggregator reporting rate consistency
Data completeness*†	Gap rate	Gap rate in data from PMU
		Gap rate in data from aggregator
	Gap size	Mean length of gaps in data from PMU
		Mean length of gaps from aggregator(s)
	Largest known gap	Largest known gap from PMU
		Largest known gap from aggregator
Characteristics of the data process path*	Dataset manageability	
	Dataset recoverability	
	Dataset reliability	
	Dataset serviceability	

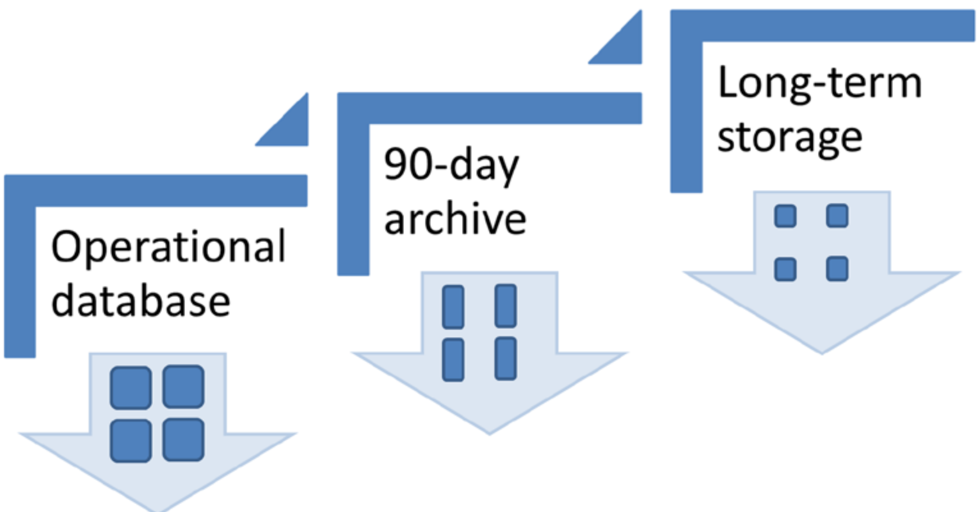
* Assess regularly – this attribute does not need to accompany every measurement, but should be weighed when determining the suitability of data for an application or study or when troubleshooting problems with application results.

† There are a number of statistics that could be used to characterize a dataset or an incoming PMU data stream; this attribute is provided as one example. Which statistics should be chosen to consider the fitness-for-use of a dataset or an incoming PMU data stream by an application must be determined by the user for the application and specific study to be performed.

Cautions regarding archived PMU data

Power system data is often archived in multiple stages – for example, data less than three months old might reside in one system and then be moved to another system after the three-month duration. The data may be relocated again after a longer period of time typically measured in years. When performing an analysis with historical data or even planning how to manage and archive data for future use, it is important to track any transformations or changes to data made during archiving. This might include compression or even deletion of repetitive measurements.

Compression rules will often vary in each of the systems. Examples include deletion of unvarying data, saving only key inflection points in the data, or truncation of significant digits beyond a certain threshold. Each successive compression or transformation may diminish the accuracy, precision and/or availability of the PMU dataset for future use, as illustrated in the figure below.



Successive archive-related data transformations and compressions may cause irretrievable changes in the quality and character of PMU data.

An additional concern is that archive storage does not necessarily require real-time time alignment of data in the way that a PDC does. However, many real-time applications DO require the time-alignment capabilities of the generalized aggregator (in this case, the PDC).

Again, the *data content* portions of *data lineage* of datasets include metadata that is typically part of a registry database like that of Table 2-4. While these attributes do not change frequently, they reflect characteristics that are necessary to properly interpret and evaluate the datasets. These data content attributes include:

- **Number of PMUs**
- **Channels provided by the PMU**[12]

[12] Here, a phasor is considered one channel because it represents one measurement (its physical input to the device is one conductor), though it could be thought of as two channels because it is a complex value. A, B, C, + seq. could be considered separate channels; however, for the purposes of this document, what matters most is that a dataset specifies clearly what channels are included rather than the exact definition of a channel.

- **Channels provided by the aggregator**, including aggregator metrics that are included in the data streams
- **PMU reporting rate** (frames per second)
- **Aggregator reporting rate** (frames per second)

2.3.2 Logical Consistency

The next overall attribute category for the datasets is logical consistency. Logical consistency refers to adherence to the structure of a dataset (Buckley1997). Unrecognized variations or inconsistencies of attributes such as measurement rates or header labeling between datasets could produce erroneous or misleading results from a synchrophasor application. PMU registries (e.g., that shown in Table 2-4) and data protocols are important tools to ensure logical consistency across synchrophasor datasets.

In the attributes below, persistence refers to an attribute remaining over time, or remaining the same over time, while consistency refers to an attribute remaining or remaining the same across an aggregator or dataset.

Attributes of logical consistency include:

- **Metrology persistence and consistency**
  - **PMU metrology persistence** – whether the measurement in question remains determined by the same method over time by the PMU
  - **Aggregator metrology persistence** – whether the measurement in question remains determined by the same method over time for each synchrophasor device contributing to the aggregator
  - **Aggregator metrology consistency** – whether the measurement in question is determined by the same method across all synchrophasor devices contributing to the aggregator
- **Header persistence and consistency**
  - **PMU header persistence** – whether the PMU header structure stays the same over time
  - **Aggregator header persistence** – whether the aggregator header structure stays the same over time
  - **Aggregator header consistency** – whether the header structure stays the same across all synchrophasor devices contributing to the aggregator
- **Data frame persistence and consistency**
  - **PMU data frame persistence** – whether the PMU data frame structure stays the same over time
  - **Aggregator data frame persistence** – whether the aggregator data frame structure stays the same over time
  - **Aggregator data frame consistency** – whether the data frame structure stays the same across all synchrophasor devices contributing to the aggregator
- **Data frame order consistency**
  - **Aggregator data frame order consistency** – whether the data frames at an aggregator are received and recorded in the correct order

- **Standards compliance persistence and consistency**[13]
    - **PMU standards compliance persistence** – whether the standards compliance of the PMU stays the same over time
    - **Aggregator standards compliance persistence** – whether the standards compliance of the aggregator stays the same over time
    - **Aggregator standards compliance consistency** – whether the standards compliance of the aggregator stays the same across all synchrophasor devices contributing to the aggregator
- **Reporting rate persistence and consistency**
    - **PMU reporting rate persistence** – whether the PMU reporting rate stays the same over time
    - **Aggregator reporting rate persistence** – whether the aggregator reporting rate stays the same over time
    - **Aggregator reporting rate consistency** – whether the reporting rate stays the same across all synchrophasor devices contributing to the aggregator

Logical consistency may be lost under the following conditions:

- Data ships from the PMU without the required headers/identifiers.
- Data is mislabeled at an aggregator. If an individual PMU changes the channels it is sampling and sending data for without the aggregator being adjusted accordingly, not only does much or all of the data for that PMU become mislabeled, later data that pass through that aggregator may also become mislabeled.[14]
- Data is duplicated at an aggregator.
- Data is shipped from the aggregator without the required headers or identifiers.
- Data is mislabeled when read into a storage system or archive for processing.
- Data is duplicated when read into a database for processing.
- Multiple data from different PMUs are given erroneous timestamps (as from incorrect GPS processing of the leap-second), so they are misaligned at the aggregator; or one set of data from a single PMU writes over another dataset because they both received the same timestamp before and after the period when the GPS times were incorrectly determined.

Some of these are also listed as sources of error under *data accuracy*. Mislabeling of data might be perceived as either, depending on how it is discovered (see text box for an example). Technically, mislabeling is a problem of logical consistency, but if the mislabeled data is fed into an application, it may look like wildly inaccurate data and produce a misleading analytical result.

---

[13] Standards compliance includes the possibility that the vendor does not build the device to a specific technical standard, or has not substantiated claims to comply with a particular standard.

[14] In the case of aggregator-to-aggregator communication, this depends on the communication protocol used. IEEE C37.118, for instance, is susceptible to mislabeling, but open source protocols such as GEP are not nearly as susceptible to this. In general, it is worth noting that modes of failure can be influenced by the communication protocol.

Examples of Mislabeling

Mislabeling appearing at the aggregator (PDC)
The data cleaning algorithms for the Data Integrity Situational Awareness Tool (DISAT) would label values for a particular data channel out of bounds if they exceeded certain thresholds of deviation from expected values. The values too far out of range would then be discarded, because DISAT has a high tolerance for gaps in the data. The developers then asked power systems engineers to review the data channels that were out of bounds for very long periods, and in most cases this made it clear that the data had been mislabeled at an aggregator, and the suspect values were within expected ranges once the correct header was known (Amidan 2014).

Mislabeling due to erroneous timestamps
Eight PMUs monitored over the leap second of June 30, 2015 had timestamps that did not correctly reflect the time of the measurement. This condition existed from between 0.9 seconds and 16 seconds (depending on the PMU model) before being corrected, so for a while there would be two sets of data with the same timestamp (Goldstein 2015).

2.3.3 Data Completeness

Data completeness refers to whether there are any gaps in the data and, if so, the size of the gaps. Data completeness attributes describe gaps not described by data lineage attributes; data lineage describes a set or subset of data in overall terms of what is intended or expected to be in the data; and data completeness describe how some of the intended data may be missing.

Attributes of data completeness include the following:

- **Gap rate** representing the number of gaps in the data per unit time, where a gap is defined as the time from when measurements stop to the time when measurements start again
  - **Gap rate in data from PMU**
  - **Gap rate in data from aggregator**
- **Mean gap size** representing the mean of the sizes of the known gaps in the data in units of time
  - **Mean length of gaps in data from PMU**
  - **Mean length of gaps from aggregator(s)**
- **Largest known gap** representing the size of the largest known gap in the data in units of time
  - **Largest known gap from PMU**
  - **Largest known gap from aggregator**

Example causes of loss of completeness include:

- A PMU ceases to function
- A functioning PMU's communications link ceases to function
- An aggregator ceases to function
- An aggregator overwrites a needed data channel
- A functioning aggregator's communications link ceases to function
- Network errors drop out some data points.

Most causes of data completeness problems arise from poor performance by a PMU, aggregator, or network communications element, and can be eliminated (for future measurements) through better management and more attentive business practices.

2.3.4 Characteristics of the Data Process Path Affecting Datasets

Because the data process path affects data quality in many ways, it is useful to describe some characteristics of the data process path, distinct from the attributes of data points, datasets, and data streams.

Certain concepts that fall under the umbrella of availability – manageability, recoverability, reliability, and serviceability – apply both to datasets and to incoming data streams. However, the exact attributes needed to describe the data will differ somewhat for static datasets versus additional availability attributes for incoming data streams. Similarly, applications may have availability requirements only needed for processing incoming streams that are not needed for processing static datasets. This section discusses availability attributes for static datasets, while availability attributes for incoming data streams are discussed in a later section.

Individual PMU availability (manageability, recoverability, reliability, and serviceability) should be well-specified by the manufacturer (presuming its power supply and supply of the signal being measured remain operational). In addition, these same characteristics of manageability, recoverability, reliability, and serviceability apply to the process of gathering datasets and moving them from the point of measurement to the point of use. Sound business practices and explicit quality of service requirements are essential for attaining and maintaining high data quality. Data process path performance requirements should be clearly articulated in designing and procuring synchrophasor system elements.

- **Dataset manageability** – the ability to create and maintain an effective quality of service that delivers the PMU dataset to the application.
- **Dataset recoverability** – the ability to re-establish delivery of the PMU dataset to the application after an interruption. For datasets, this recoverability can extend beyond just the network communications path being restored. It can also include retrieval from remote data archives manually (e.g., fault data recorders) or assembling from off-site archives of the same data. As a result, it is up to the organization to decide on what timeframe must be met for satisfactory recoverability based on the resources available.
- **Dataset reliability** – the ability to deliver the PMU dataset to the application at a specified message rate for a stated period.
- **Dataset serviceability –** the ability to identify the existence of problems in the PMU dataset delivery, diagnose their cause(s), and repair those problems. An application's requirements for serviceability are largely covered under its requirements for reliability. However, because high-reliability communications networks are not cost-free, an organization may need to define both reliability and serviceability requirements for its PMU data stream that fit its resources.

2.4 Illustrating Dataset Attributes

Due to the level of complexity of the model proposed for describing data quality, it may be easier to describe dataset problems using a common template and color scheme to illustrate many of the attributes and flaws that characterize single data points and datasets.

Figure 2-3 gives the legend for all diagrams in this section. Later diagrams use these colors and shading consistently, but some demonstrate different time durations and PMU reporting rates.

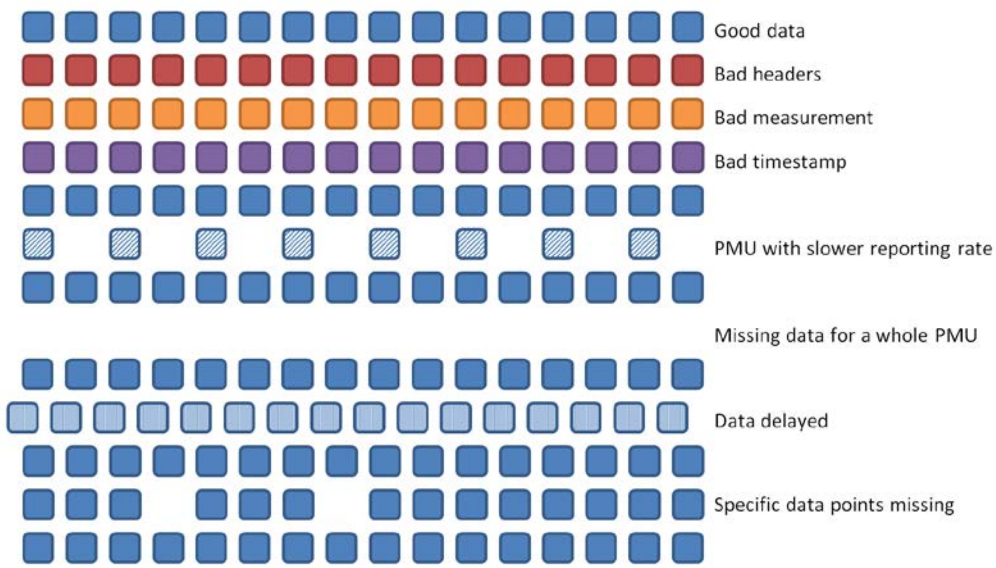


Figure 2-3: Schema for Example Diagrams of the Various Metrics and Items that Can Affect Them

2.4.1 Attributes of Single Data Points

Attributes of quality and fitness-for-use associated with single data points are typically tied to the measurement aspects of the synchrophasor data. Many specific metrics of data weakness (e.g., measurement specifiers, measurement accuracy, and attribute accuracy) fold into a larger category of "bad measurements." Figure 2-4 shows how bad measurement sets can look in the dataset, using dark blue dots to show good data and orange dots to show data points with lower data quality.

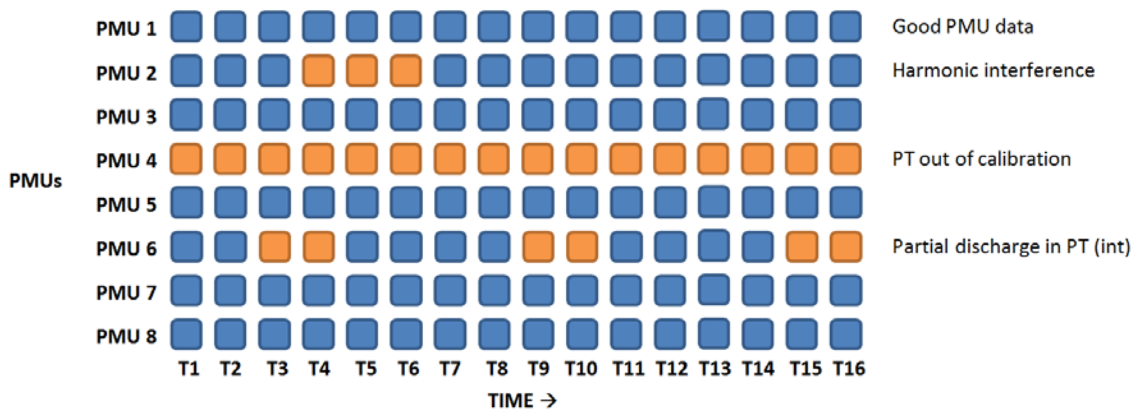


Figure 2-4: Bad Measurements Example

Figure 2-4 shows three types of bad measurements:

- The first is harmonic interference, which appears in PMU 2. PMU 2 is a P-class PMU, and the harmonic interference is caused intermittently by another device near the PMU, so only a few data reports are corrupted.
- The second is a PT being out of calibration, which appears in PMU 4. In this case, the PT feeding into the PMU was not calibrated properly, so the overall measurement accuracy is affected. As such, the entire dataset reported from PMU 4 is questionable. Any applications using this data

should be evaluated to determine how this inaccurate measurement how that might affect analytical results.

- The last is an intermittent bad measurement, which appears in PMU 6. PMU 6 has an intermittent bad measurement associated with an intermittent partial discharge from a failing PT, causing bad measurements in the dataset.[15]

2.4.2 Attributes of Datasets

Attributes of datasets encompass larger characteristics and measurements of the synchrophasor data. This can include information on the data grouping as a whole, such as if the same reporting rate and methods of measurement are present throughout the full set.

2.4.2.1 Metrology Persistence

Metrology persistence throughout the dataset can be influenced by a variety of factors. Turns-ratio changes on instrumentation transformers, relocation of instrument transformers, and even simple scalar changes can affect whether how the measurements are taken remain the same over time. Figure 2-5 shows an example with a PMU that was reconfigured.

Figure 2-5 represents data collected from a single PMU. In this particular instance, a secondary set of instrumentation transformers was removed, causing three measurements to no longer be available from that PMU. If a particular application were utilizing these channels (Voltage 2, Current 2, or Power 2), that application might either fail, or begin using incorrect channels and reporting erroneous results.

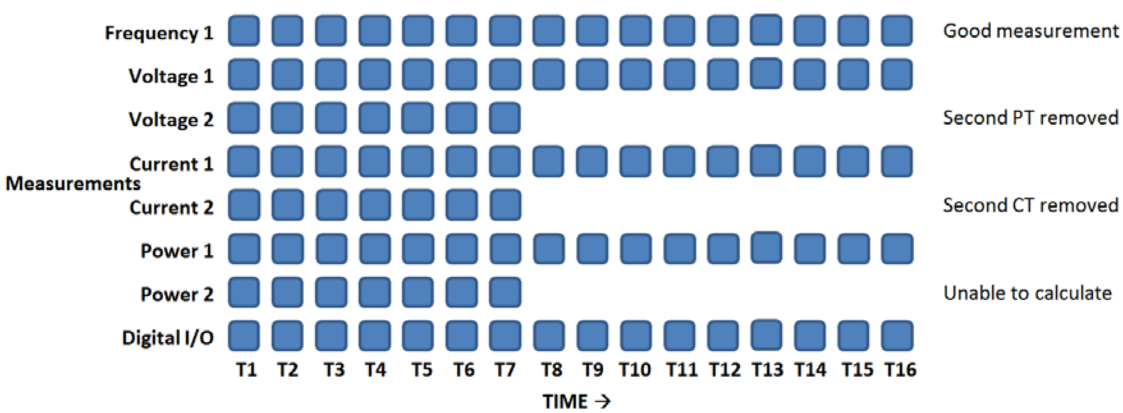


Figure 2-5: Measurement Persistence Example

2.4.2.2 Header Persistence and Consistency

Poor initial configuration or an equipment update may compromise the header information associated with a synchrophasor device or aggregator. Figure 2-6 shows three scenarios that may yield inconsistent or incorrect header information, as indicated by the red squares.

[15] If the measurements associated with this channel start to show an odd or intermittent pattern, an alert engineer would look for the causes of those odd measurements – in other words, are the anomalous measurements revealing bad measurements or accurate measurements of bad grid conditions? In the case of a failing instrument transformer, even bad measurements can serve a valuable purpose.

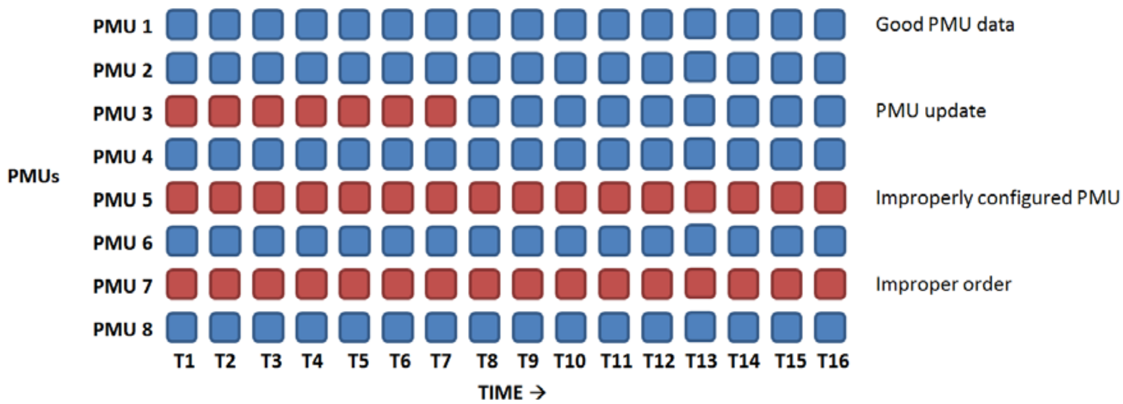


Figure 2-6: Header Persistence and Consistency Example

Most of the data shown in Figure 2-6 are good data. The following three header concerns are present:

- PMU 3 was upgraded part way through this period, changing the headers after the upgrade, so there is no persistence for PMU 3 header information over all time in this dataset. This may not cause any issues for the end-use application, but is a case the application designer and tester must include.
- PMU 5 represents a PMU that is not configured properly, so the header information is incorrect. While the header information is persistent over time for all data coming from PMU 5, it is not consistent with the information encoded in the rest of the dataset. Depending on the application, this may or may not cause issues, notably if the header information is being utilized.
- PMU 7 data represents inconsistent header information caused by aggregator configuration problems. Once again, the information is persistent over time for all data coming from PMU 7, but in this case it is incorrect for the underlying data. In this particular case, a PDC is operating on an assumed data input order, but a PMU has dropped off. As such, the header information associated with PMU 7 is actually for a different PMU, and is no longer valid.

2.4.2.3 Data Frame Persistence and Consistency

Data frame persistence and consistency is very similar to header consistency. Figure 2-7 shows the result of an aggregator misconfiguration. The aggregator improperly handles the ordering of data, so a data gap in PMU 1 shifts PMUs 2 through 5 up one position. For the data frames associated with T5 through T9, PMUs 2 through 4 are actually reporting as PMUs 1 through 3. The result is that for those four time periods, the data in that frame is not consistent with the rest of the dataset. If the application is only parsing the frame layout at the beginning of this period, it would not correctly handle this transition and would potentially use incorrect data.

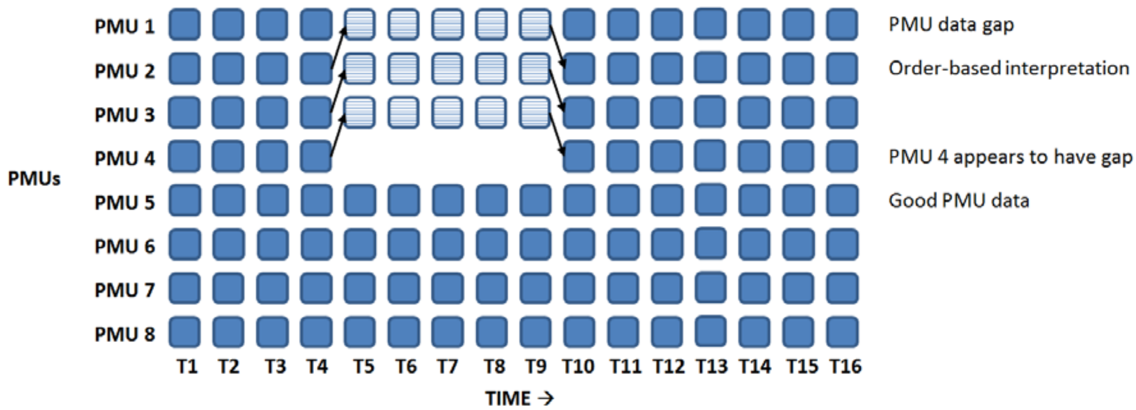


Figure 2-7: Data Frame Persistence and Consistency Issues Example

2.4.2.4 Data Frame Order Consistency

Communications network issues can cause data frames to arrive out of order at an aggregator, possibly causing data frames to be stored in the wrong order in the dataset. Figure 2-8 shows an example in which one frame is delayed, so that several frames are out of order.

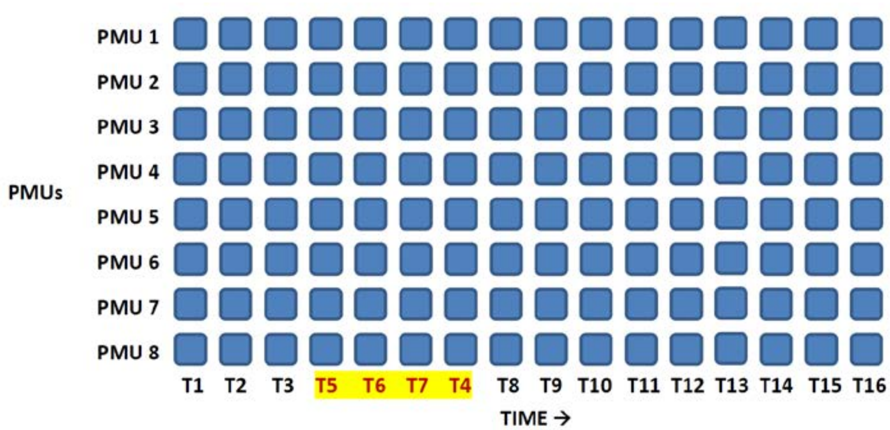


Figure 2-8: Data Frame Order Consistency Issues Example

2.4.2.5 Reporting Rate Persistence and Consistency

Reporting rate persistence and consistency for both PMU data and aggregator data can be a challenge for the end-use application. If a PMU is upgraded to a newer version, the internal measurements may change, resulting in different reporting capabilities from within the PMU – the reporting rate might still look the same, but the answers would now be obtained differently. Figure 2-9 shows an example of an updated PMU with a higher measurement sampling rate and an associated higher reporting rate. As a result, the associated metrics with the measurements, such as the accuracy and precision, can be significantly different from T7 onward. If the end-use application is sensitive to this change, the results may be affected. Figure 2-10 shows examples of other equipment updates or functionality changes impacting reporting rate persistence and consistency for a set of PMUs reporting to an aggregator.

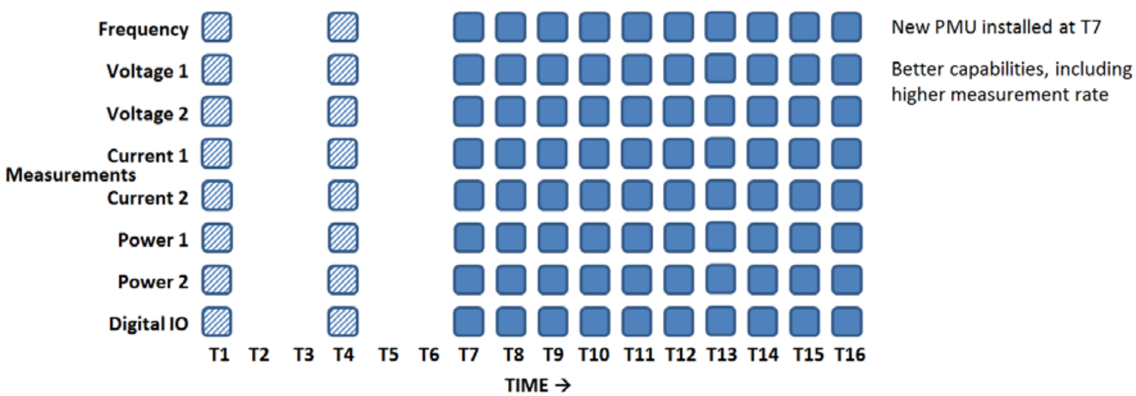


Figure 2-9: Measurement Rate Persistence Example For a Single PMU

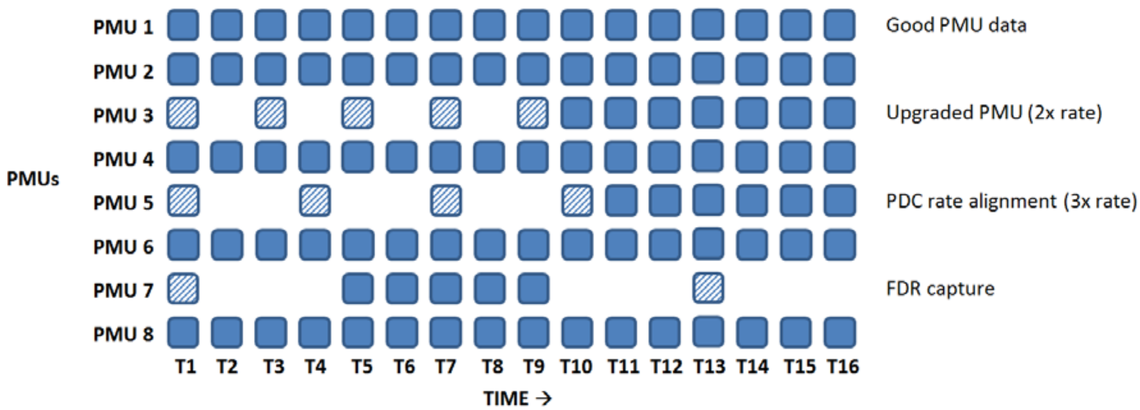


Figure 2-10: Reporting Rate Persistence and Consistency Example

Most of the datasets shown in Figure 2-10 are good data from the reporting rate persistence perspective. This example shows three other reporting rate persistence and consistency concerns:

- In PMU 3, the PMU was replaced with a newer, faster-reporting PMU. As a result, the reporting rate is not persistent across the data, so any applications using this data would need to accommodate this feature.
- PMU 5 is similar, except its reporting rate was adjusted to align with an aggregator (PDC) later in the system. The reporting rate was updated from a much slower rate when the downstream PDC was upgraded, because it can now handle the increased reporting rate.
- PMU 7 data represent a case where the synchrophasor device is not solely a PMU. PMU 7 is actually a device with fault disturbance recording capability. After the triggering event occurs, the device performs at a much higher reporting rate for a few seconds.

2.5 Attributes of an incoming PMU data stream

While the attributes of datasets describe the fitness-for-use of a given dataset once gathered, there are additional attributes that must be considered for an application that processes an incoming PMU data stream. As for static datasets, both from concentrators and from individual PMUs, characteristics of the data process path and the practices that determine those characteristics can have a profound impact upon a streaming dataset, even though that impact can differ for batch processing a static dataset versus processing an incoming PMU data stream.

Data stream availability is an umbrella term for whether an incoming PMU data stream feeds data into an application with sufficient timeliness and dependability for the application to deliver operationally useful results.

A PMU data stream is a set of data points which arrive over a set of time intervals, or more formally as any ordered pair *(s,Δ)* such that *s* is a sequence of tuples and *Δ* is sequence of positive real-time intervals. The first set of incoming PMU data stream attributes characterize the data process path that delivers the data from the PMU to the application; the second set of incoming PMU data stream attributes characterize the availability of the incoming PMU data stream. Table 2-6 summarizes the categories of attributes and metrics associated with an incoming PMU data stream.

Table 2-6: Incoming PMU Data Stream Attributes and Metrics

Characteristics of the data process path*	Data stream manageability
	Data stream recoverability
	Data stream reliability
	Data stream serviceability
Data stream availability*†	Message rate
	Message rate type
	Message arrival order correctness
	Message delivery time
	Expected drop-out rate†
	Expected drop-out size†
	Message continuity period†

* Assess regularly – this attribute does not need to accompany every measurement, but should be weighed when determining the suitability of data for an application or study or when troubleshooting problems with application results.

† There are a number of statistics that could be used to characterize a dataset or an incoming PMU data stream; this attribute is provided as one example. Which statistics should be chosen to consider the fitness-for-use of a dataset or an incoming PMU data stream by an application must be determined by the user for the application and specific study to be performed.

2.5.1 Characteristics of the Data Process Path Affecting Data Streams

As discussed in an earlier section, certain concepts that fall under the umbrella of availability apply both to datasets and to incoming data streams (i.e., manageability, recoverability, reliability, and serviceability). However, applications that process incoming data streams have availability requirements that are not needed for processing static datasets.

Network availability is often the dominating factor affecting whether data stream availability will be satisfactory. Sound business practices and explicit quality of service requirements are essential for attaining and maintaining high data quality.

The following characteristics and needs should be understood and specified before procuring any element of the synchrophasor device and data process path. All of these attributes contain the issues associated with the dataset attributes category in a previous section, with additional considerations associated with the real-time constraints associated with an incoming PMU data stream.

- **Data stream manageability** – the ability to create and maintain an effective quality of service that delivers the PMU data stream to the application.
- **Data stream recoverability** – the ability to re-establish the PMU data stream to the application after an interruption. Requirements an application may have for recoverability are largely covered under its requirements for reliability. However, for an organization to determine whether it can meet an application's reliability requirements, that organization may need to define some recoverability requirements for its PMU data stream that fit that organization's resources.

- **Data stream reliability** – the ability to deliver the PMU stream to the application at a specified message rate for a stated period.
- **Data stream serviceability** – the ability to determine the existence of problems in the PMU data stream delivery, diagnose their cause(s), and repair those problems. An application's requirements for serviceability are largely covered under its requirements for reliability. However, because high-reliability communications networks are not cost-free, an organization may need to define both reliability and serviceability requirements for its PMU data stream that fit the organization's resources.

2.5.2 Data Stream Availability

The attributes of data stream availability include the following:

- **Message rate** – the expected number of message reports per second. Note that this is not necessarily the same as the measurement rate defined under data quality. Ideally, the measurement rate and the message rate would be the same number and always constant, but realities of measurement and network transport mean that they can – and do – differ.
- **Message rate type** – whether or not the message rate is expected to be fixed or variable.
- **Message arrival order correctness** – the messages of a data stream arrive in the correct order (i.e., the order of increasing timestamps).
- **Measurement delivery time** – the time between the timestamp on a measurement and its arrival at the application.
- **Expected drop-out rate** – the rate at which messages may be lost from the data stream, but the data stream is still considered live and operational.
- **Expected drop-out size** – the number of messages in a row that may be missing while the data stream is still considered live and operational.
- **Message continuity period** – a period of time for which the message rate remains as expected, with no more drop-offs than the *expected drop-off rate* and not drop-off exceeding the *expected drop-off size*.

2.5.3 Illustrating Incoming PMU Data Streams

Some characteristics can take on slightly different meaning, depending on the context. If a dataset is off-line rather than being streamed, similar-looking attributes may have different meaning.

2.5.3.1 Gap Rate and Data Drop-out Rate

Gap rate and data drop-out rate can look very similar. Gap rate primarily deals with whether the information exists in any form and is obtainable, whereas drop-out rate is associated with timely or successful receipt of streamed data. A drop-out in a stream may result in a gap, but that gap may be fillable with another database or some form of local information. Both gaps and drop-outs represent a loss of information, but drop-outs are associated with loss of that information in a near-term sense, whereas gaps indicate that data does not exist at all.

Figure 2-11 shows an example of gaps in the data. Whether due to network drop-out, database storage errors, or even just missing data from the PMU itself, all of the channels are experiencing missing data. It is also possible in some cases that significant network delay effectively turns late data into missing data because of time limits set within the receiving PDC. This cumulative probability distribution of data loss is called 'data drop-out' and characterized using the metrics drop-out size (gap size) and drop-out rate (gap rate). The sizes of the largest gaps and drop-outs are given on the right, ranging from 1 to 4 time units. If an application can only handle a 1-datapoint gap, it may mis-perform given the data coming in from PMUs 1, 4, 5, and 7. This could cause a failure of the application (e.g., a state estimator could fail to converge) or just an indication the results are not as trustworthy as expected.

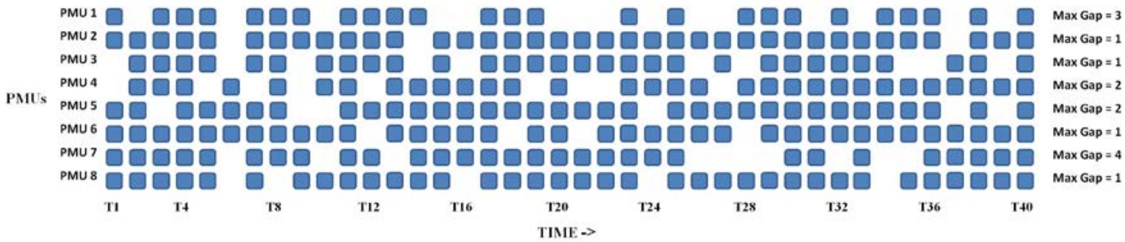


Figure 2-11: Data Drop-Outs Associated with a Network with Random Data Loss Probability

Another metric associated with Figure 2-11 is the gap rate. While the maximum gap of PMU 8 is only one data packet, those missing points occur five times in the 40 report intervals shown. If an application can only handle a gap of 1 data point in every 40 reports, it would give undesirable or unexpected results for this PMU.

Clearly, each additional aggregator and network link between the PMU and the application increases the probability of data drop-outs. Figure 2-12 illustrates the case where an additional data aggregator is introduced into the data stream shown in Figure 2-11, causing additional data drop-outs.

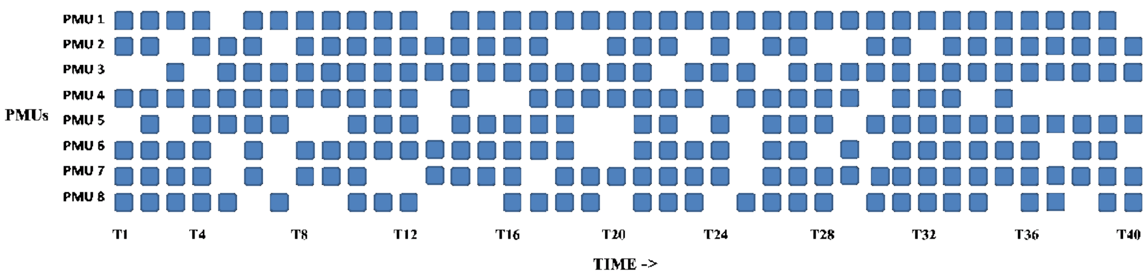


Figure 2-12: Data Drop-outs Associated with a Network with Random Loss Probability, Coupled with a Data Aggregator Introducing Random Processing Delays

The specific drop-out rates (number of drop-outs) and sizes (number of drop-outs in a row) for the datasets illustrated in Figure 2-11 and Figure 2-12 are plotted in Figure 2-13 and Figure 2-14, respectively. The analyses in Figure 2-13 and Figure 2-14 consider an application that requires 10 sequential measurements in order to operate, and achieves this sample selection by moving a window spanning the 10 most recent measurements. The drop-out rate and size, as perceived by the application, appear to be random.

Consider the plots in Figure 2-13. If an application requires a drop-out rate of less than half the measurements in any given sample window in order to function (drop-out rate < 0.5), then the data streams from PMUs 1 and 7 violate this requirement at time indices 13 and 22, respectively. The data stream from both PMUs 1 and 7 return to specification at time indices 25 and 28, respectively.

If the application requires that the rolling maximum drop-out size to be less than 4, then PMU 7 violates this requirement at time index 19 and returns to specification at time index 22. Tolerance to this kind of periodic “bursty” loss of data is critical to applications that require real-time Ethernet-transported measurements. Assuming that network congestion is controlled and that the network is physically operational, data drop-outs typically occur in short close bursts.

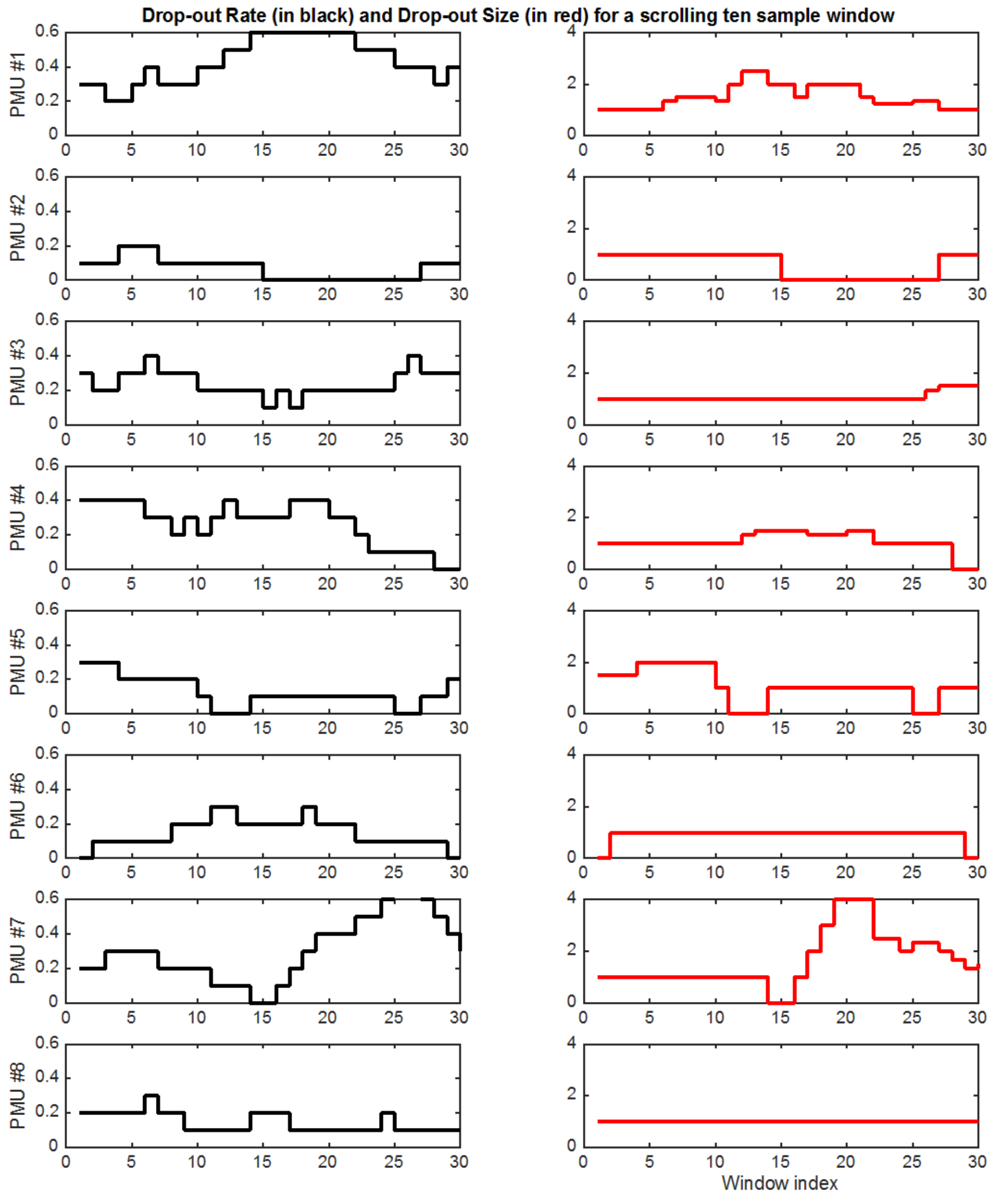


Figure 2-13: Data drop-out rate (in black) and Drop-out size (in red) for a scrolling ten sample window (ref: Figure 2-11)

Now consider the graphs in Figure 2-14 showing drop-out rate and size for the same network and data streams, but passing through data aggregators. Data aggregators provide some buffer capacity to the network link. In many cases this buffer helps reduce the drop-out rate and size by recovering and reordering delayed measurements. However, since the aggregator has finite buffer capacity and introduces computational overhead, it may also introduce its own stochasticity into the incoming data stream. Figure 2-14 illustrates one such case where the introduction of an aggregator helps improve the data stream's ability to meet the application specification; drop-out size less than 4 and rate less than 0.5. However, the increasing drop-out rate in the case of PMU data stream 4 may indicate buffer over-runs. Also, the baseline (average across all 30 sample selections) drop-out rate is higher for most PMUs in Figure 2-13. When compared to Figure 2-14, this illustrates the uniformly distributed probability of drop-outs due to busy processors on the aggregator when handling multiple network connections.

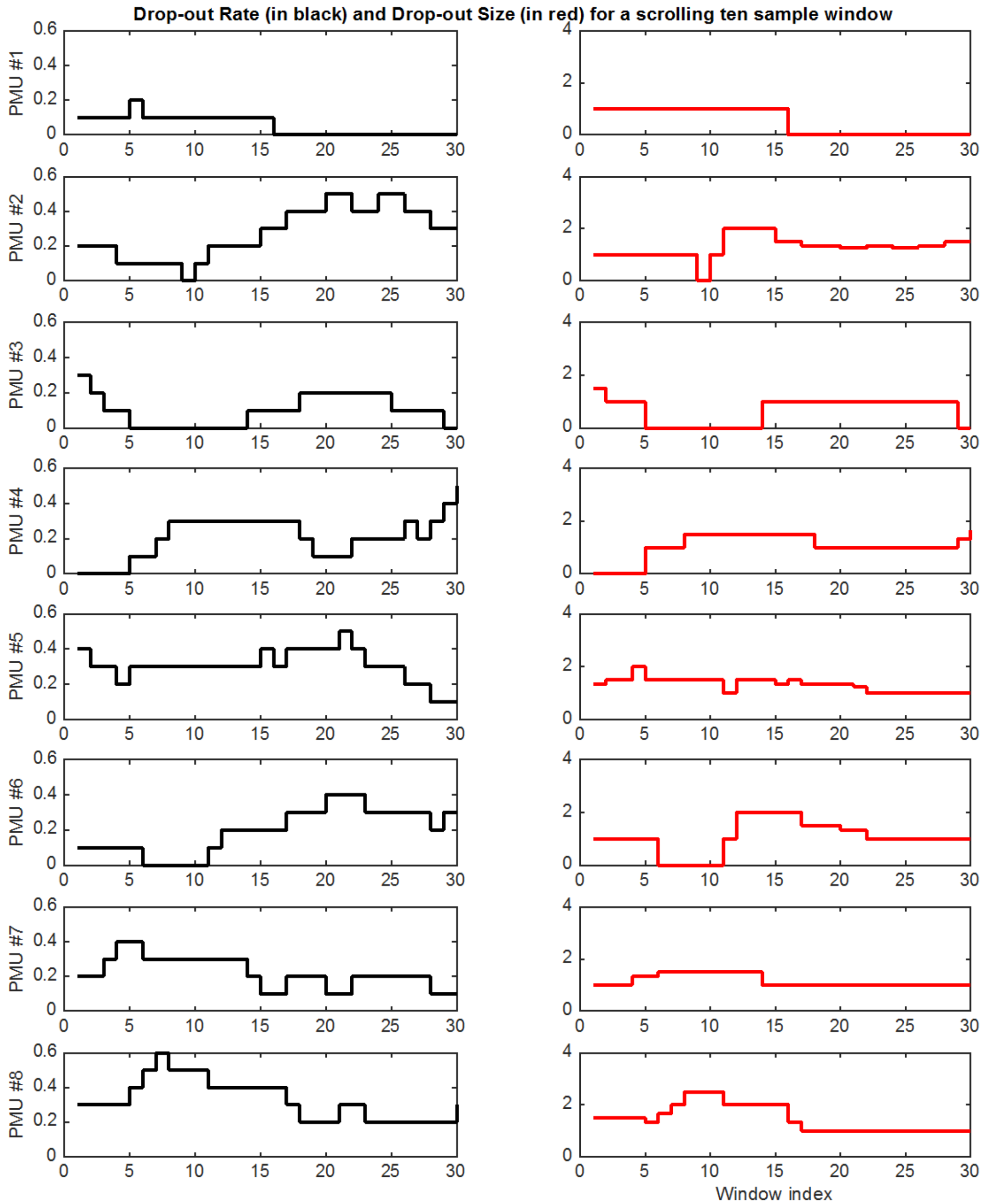


Figure 2-14: Data Drop-out Rate (in Black) and Drop-out Size (in Red) for a Scrolling Ten Sample Window of Data Passing Through a Network and Aggregators (ref: Figure 2-12)

2.5.3.2 Network Latency

Network latency is a critical facet of data streams. After a synchrophasor device has obtained a measurement, it must communicate that with the application. Near-real-time applications may need to receive the PMU data within a few milliseconds of its measurement, and network latency will affect the delivery speed.

Figure 2-15 shows a simple example of how network latency can affect when the data arrives at the application. If latency is too high, it may either violate the real-time constraints of the application, or cause the network to skip or drop that data packet (and many others).

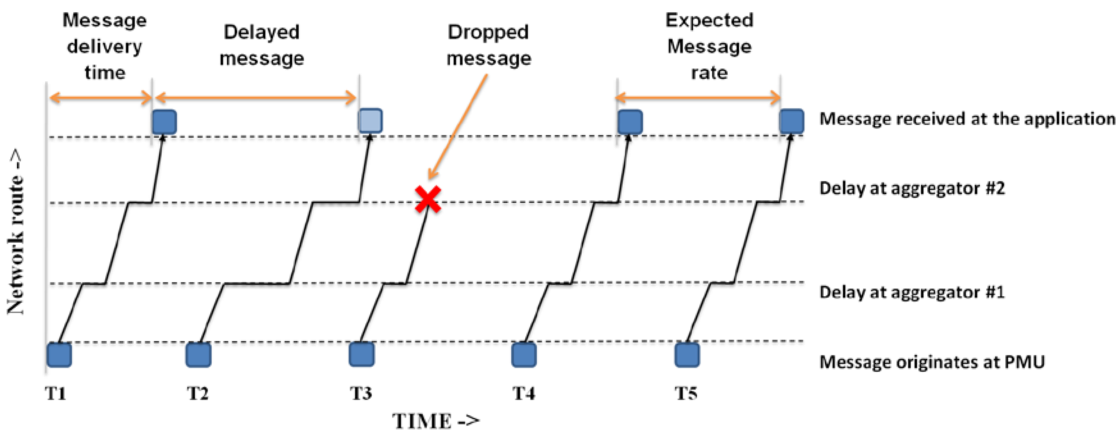


Figure 2-15: Network Latency Diagram

Figure 2-16 shows a more detailed example of network latency on the system. Data associated with PMU 1 through 5 is still making it to the endpoint intact. However, data from PMUs 3, 4 and 5 are being received at the same packet rate PMU 1 and 2 are providing. Intermittent reports from PMU 3 are delayed in their transmission, and the data from PMU 4 is being delayed due to routing congestion. PMU 5 is experiencing routing errors, which may be significantly delaying the receipt of its data packets. If the application is extremely sensitive to the consistent arrival of data packets, the information from PMUs 3, 4, and 5 may be missed, or may delay the update of the algorithm. The application designer will need to either account for this delay, or determine the threshold after which the data is simply ignored.

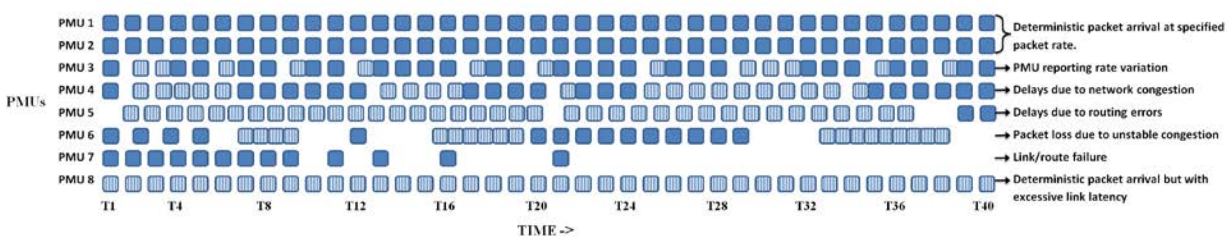


Figure 2-16: Network Delays Example

In Figure 2-16, PMUs 6 and 7 represent cases of network latency where data is lost, either due to unstable congestion or a link or route failure. This can be caused by factors such as excessive congestion causing a router to drop the individual messages, or even a complete failure of a device in the communications path. If the application fails to recognize this missing data, or if it is critical information, the application could produce erroneous or skewed results.

PMU 8 in Figure 2-16 is still arriving at the proper, specified rate, but has a very consistent latency. As such, even though the data appears aligned with the expected rate of PMUs 1 and 2, it may actually be delivering data several time intervals behind the other PMUs. If the application is not properly recognizing and aligning these delays, or if the delay violates the application's real-time requirements, the late data may not be included or may affect the overall performance of the application.

Aggregation of data can introduce delays

As data packets get larger as they are aggregated, they may grow large enough that they have to be fragmented into smaller packets in order to be transmitted further along the data process path. This can happen for more than one transport protocol and at more than one location in the data process path, and can introduce delays as well as dropouts and data corruption.

In addition, aggregators such as PDCs at the substation and further along the data process path may have to wait to output any given frame until all the relevant reporting to the aggregator have sent values for that frame.

2.6 Summary of Data Attributes

Table 2-7 summarizes the broad categories and detailed data attributes. Appendix A provides these attributes in further detail, including the metric or specification to measure the attribute, relevant application requirements, and the impacts of gaps or problems with data relevant to each attribute.

Table 2-7: Summary of Data Attributes

<table>
<tr><th colspan="3">Attributes of single data points</th></tr>
<tr><td colspan="2" rowspan="5">Measurement specifiers</td><td>Signal being measured</td></tr>
<tr><td>Measured result</td></tr>
<tr><td>Standard units</td></tr>
<tr><td>Precision</td></tr>
<tr><td>Confidence interval</td></tr>
<tr><td colspan="2" rowspan="2">Measurement accuracy</td><td>Source influence:
• Noise in the power system
• Out-of-band interfering signals
• Harmonics
• Instrument channel errors
• GPS loss, spoofing, or meaconing</td></tr>
<tr><td>Induced error (created by the PMU estimation process):
• Synchrophasor TVE
• Synchrophasor magnitude error
• Synchrophasor phase error
• Frequency error
• ROCOF error</td></tr>
<tr><td colspan="2" rowspan="3">Attribute accuracy</td><td>Temporal accuracy (timestamp matches time of measurement)</td></tr>
<tr><td>Geospatial accuracy (coded PMU location matches actual location)</td></tr>
<tr><td>Topological accuracy (coded PMU topological location matches location in actual power system topology)</td></tr>
<tr><td rowspan="7">Data lineage
(i.e., metadata)*</td><td rowspan="5">Data source</td><td>PMU type</td></tr>
<tr><td>PMU standard followed</td></tr>
<tr><td>PMU model, firmware version, configuration settings</td></tr>
<tr><td>PMU-supplied data headers</td></tr>
<tr><td>Aggregator-supplied data headers</td></tr>
<tr><td rowspan="2">Data coverage</td><td>PMU geospatial location</td></tr>
<tr><td>PMU topological location</td></tr>
</table>

Table 2-7: Summary of Data Attributes (contd)

Attributes of datasets		
Data lineage (i.e., metadata)*	Data transformation methods	Transformations applied to the data at the PMU
		Transformations applied to the data at an aggregator
		Transformations applied to the data during archiving
	Data coverage	Aggregator geospatial coverage
		Aggregator topological coverage
		PMU temporal coverage (time in operation)
		Aggregator temporal coverage (time in operation)
	Data content	Number of PMUs
		Channels provided by the PMU
		Channels provided by the aggregator
		PMU reporting rate
		Aggregator reporting rate
Logical consistency*	Metrology persistence and consistency	PMU metrology persistence
		Aggregator metrology persistence
		Aggregator metrology consistency
	Header persistence and consistency	PMU header persistence
		Aggregator header persistence
		Aggregator header consistency
	Data frame persistence and consistency	PMU data frame persistence
		Aggregator data frame persistence
		Aggregator data frame consistency
	Data frame order consistency	Aggregator data frame order consistency
	Standards compliance persistence and consistency	PMU standards compliance persistence
		Aggregator standards compliance persistence
		Aggregator standards compliance consistency
	Reporting rate persistence and consistency	PMU reporting rate persistence
		Aggregator reporting rate persistence
		Aggregator reporting rate consistency
Data completeness*†	Gap rate	Gap rate in data from PMU
		Gap rate in data from aggregator
	Gap size	Mean length of gaps in data from PMU
		Mean length of gaps from aggregator(s)
	Largest known gap	Largest known gap from PMU
		Largest known gap from aggregator
Characteristics of the data process path*	Dataset manageability	
	Dataset recoverability	
	Dataset reliability	
	Dataset serviceability	

* Assess regularly – this attribute does not need to accompany every measurement, but should be weighed when determining the suitability of data for an application or study or when troubleshooting problems with application results.

† There are a number of statistics that could be used to characterize a dataset or an incoming PMU data stream; this attribute is provided as one example. Which statistics should be chosen to consider the fitness-for-use of a dataset or an incoming PMU data stream by an application must be determined by the user for the application and specific study to be performed.

Table 2-7: Summary of Data Attributes (contd)

Attributes of an incoming PMU data stream	
Characteristics of the data process path*	Data stream manageability
	Data stream recoverability
	Data stream reliability
	Data stream serviceability
Data stream availability*†	Message rate
	Message rate type
	Message arrival order correctness
	Message delivery time
	Expected drop-out rate†
	Expected drop-out size†
	Message continuity period†

* Assess regularly – this attribute does not need to accompany every measurement, but should be weighed when determining the suitability of data for an application or study or when troubleshooting problems with application results.

† There are a number of statistics that could be used to characterize a dataset or an incoming PMU data stream; this attribute is provided as one example. Which statistics should be chosen to consider the fitness-for-use of a dataset or an incoming PMU data stream by an application must be determined by the user for the application and specific study to be performed.

3.0 Looking at Data from the Application Requirements Viewpoint

Section 2 looked at the data accuracy, quality, and availability attributes of data on its own. This section looks at data from the point of view of the application that has been designed to use synchrophasor data to produce a high-quality, informative analytical result. An application that receives accurate, high-quality, and highly available data is likely to produce more informative and usable analytical answers; an application that receives data that is flawed by inaccuracies, gaps, and delays is likely to produce answers caveated by greater uncertainty or fail to produce answers at all.

Fitness-for-use means that the data provides sufficient information and content for an application to provide meaningful and trustworthy results; if an application produces questionable results because of weaknesses in the incoming data, then the data are not fit for use.

Some of the earliest work on application-based data requirements was developed by the Data & Network Management Task Team (DNMTT) of the North American Synchrophasor Initiative. The DNMTT proposed Application Service Classes and example service requirements for each class,[16] as shown in Table 3-1.

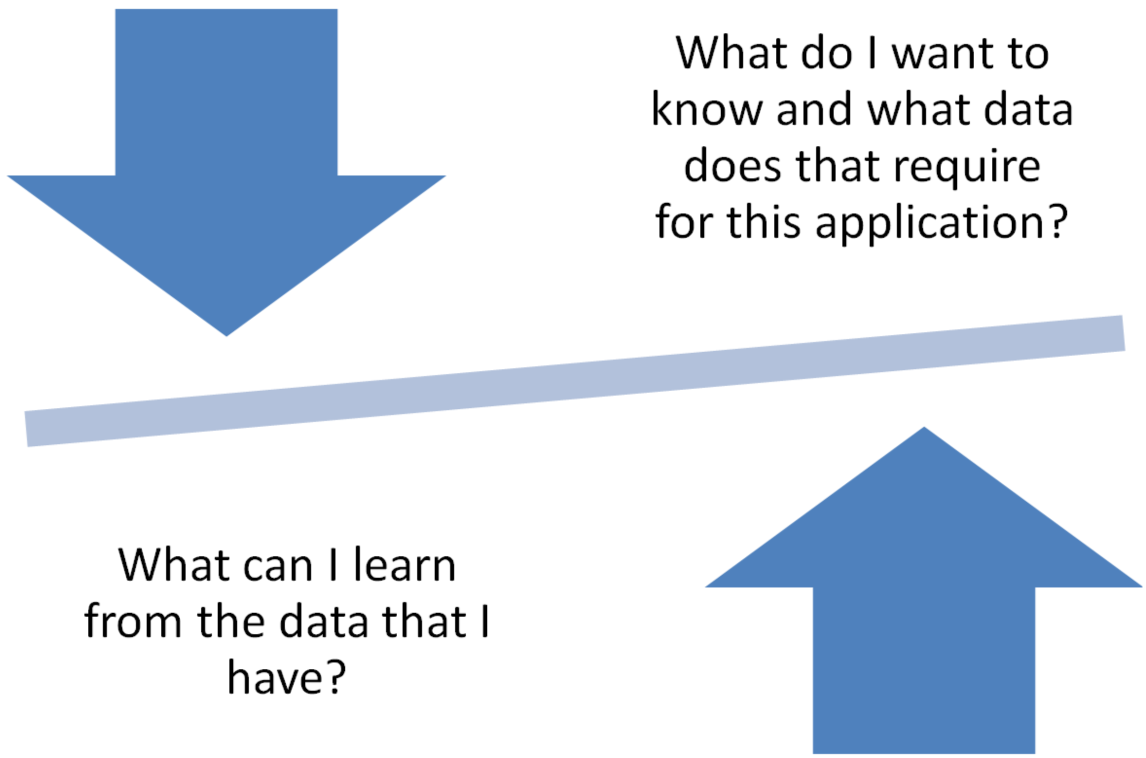
Table 3-1: NASPI Application Service Classes and Data Requirements

Application class	Application description	Data rate	Required latency
A	Feedback control	Fast	Fast
B	Open loop control	Slow	Medium
C	Post-event analysis	Fast	Slow
D	Visualization	Slow	Medium
E	Research & experimental	Varies	Varies

[16] Dagle, Jeff, "North American SynchroPhasor Initiative – An Update of Progress," Proceedings of the 42nd Hawaii International Conference on System Sciences, 2009, p. 4.

Evaluating data attributes

Which data attributes are most important depends on the problem of interest and the application designed to address it. In most cases, the analyst wants to perform a specific analysis, identifies the data needed to perform that analysis, and determines whether the data available are adequate for that purpose. In some cases, the analyst will receive a new dataset and look for what knowledge can be gained, taking the quality of the available data as a given. This dichotomy is shown below.



Which data attributes are most relevant depends on the problem at hand.

These data classes were initially developed to facilitate network design. But today, with greater synchrophasor application availability and network experience, it is clear that application requirements include data availability and integrity assurances, as well as data reporting (sampling) rates and delivery speeds.

A more recent example of application-based data requirements is shown in Table 3-2, which reflects the apparent performance expectations of most of the current and evolving synchrophasor analytical tools as well as some of the key measurement parameters for each application. Several of these parameters map directly to combinations of the data attributes discussed in Section 2:

- **Message rate** – A combination of attributes reporting rate, reporting rate persistence, and reporting rate consistency
- **Maximum measurement transfer time** – message delivery time
- **Time window** – message continuity period

- **Data loss sensitivity** – A combination of the attributes gap rate, gap size, largest known gap, dataset recoverability, expected drop-out rate, and expected drop-out size

Many of the PMU type and measurement parameters shown in Table 3-2 can be described by the data lineage attributes described in Section 2.

The discussion below does not offer specific data requirements for specific applications; that is a task to be undertaken after the synchrophasor community has reached agreement on data quality, accuracy, and availability terms. But this discussion, starting with Table 3-3, does review some of the more important data attributes and offers some thoughts and questions about how different applications might perform if the incoming data have too many flaws in that attribute.

3.1 Requirements for Data Accuracy

An application may require that measurements arrive at the application with a specified level of accuracy. For sensitive applications, this could require that the measurement algorithms used by the contributing PMUs match. An application may also have needs for accuracy of attributes such as the timestamp associated with the measurement or the exact location of the PMU within the grid topology.

3.2 Requirements for Data Content and Data Coverage

Applications will have some minimum requirements as to the content and coverage of data, but a requirement relevant to one application may be meaningless to another.

An application may require a **minimum geospatial span** or a **minimum topological span** to produce meaningful results. How those spans are defined needs to make sense for both the application itself and for the features of each service area the application will analyze.

Table 3-2: NASPI PMU Application Data Requirements

Phasor Measurement Unit Application Data Requirements											
	PMU Measurement Parameters					Delay/Quality Parameters				Other Information	
Application	*Amplitude, Angle, or Frequency Precision (p.u., degrees, mHz)*	*Amplitude, Angle, or Frequency Accuracy (%, absolute values)*	*ROCOF (Hz/s)*	*Frequency Range (Hz)*	*Time Accuracy (μs)*	*Measurement Transfer Time (ms)*	*Message Rate (Reports/sec)*	*Time Window (sec)*	*Data Loss Sensitivity (Reports or ms)*	*Performance Class (M/P/X/N)*	*Tools /Platforms*
Small-Signal Stability Monitoring	0.5 degrees 0.01 Hz	TVE	STD	0.1 - 1.0 Hz	STD	50 ms	60 Reports/sec	600 seconds	10000 ms	M	EPG RTDMS, Allstom eTerra Vision
Voltage Stability Monitoring/Assessment	0.01 p.u. mag 0.5 degrees	TVE	STD	0.1 - 10.0 Hz	STD	500 ms	30 Reports/sec	300 seconds	10000 ms	X	EPG RTDMS, Allstom eTerra Vision
Thermal Monitoring (Overload)	0.5 degrees 0.1 p.u. mag	TVE	STD	0 - 0.2 Hz	STD	1000 ms	1 Report/sec	300 seconds	30 Reports	X	
Frequency Stability/Islanding	0.5 degrees 0.01 Hz	TVE	STD	1.0 - 30.0 Hz	STD	50 ms	60 Reports/sec	5 seconds	1 Report	P	
Remedial Action Schemes: Automatic Arming	0.01 p.u. mag 0.5 degrees 0.01 Hz	TVE	STD	0.02 - 30.0 Hz	STD	20 ms	1 Report/sec	300 seconds	1 Report	P	
Remedial Action Schemes: Event Detection	0.01 p.u. mag 0.5 degrees 0.01 Hz	TVE	STD	0.02 - 30.0 Hz	STD	20 ms	60-120 Reports/sec	300 seconds	1 Report	P	
Out of step protection	0.5 degrees 0.01 Hz	TVE	STD	5.0 - 30.0 Hz	STD	10 ms	60 Reports/sec	5 seconds	0 Reports	P	
Short-term stability control	0.01 p.u. mag 0.5 degrees 0.01 Hz	TVE	STD	0.5 - 30.0 Hz	STD	16 ms	60 Reports/sec	60 seconds	10 ms	P	
Long-term stability control	0.01 p.u. mag 0.5 degrees	TVE	STD	0 - 10.0 Hz	STD	1000 ms	30 Reports/sec	600 seconds	1000 ms	X	
FACTS feedback control, Smart switch-able networks	0.01 p.u. mag 0.5 degrees 0.01 Hz	TVE	STD	1.0 - 30.0 Hz	STD	16 ms	60 Reports/sec	300 seconds	50 ms	X	
State Estimation	0.5 degrees 0.01 Hz	TVE	STD	0 - 1.0 Hz	STD	1000 ms	5 Reports/sec	300 seconds	1000 ms	M	

Table 3-2**: (contd)**

Phasor Measurement Unit Application Data Requirements											
	PMU Measurement Parameters					Delay/Quality Parameters				Other Information	
Application	*Amplitude, Angle, or Frequency Precision (p.u., degrees, mHz)*	*Amplitude, Angle, or Frequency Accuracy (%, absolute values)*	*ROCOF (Hz/s)*	*Frequency Range (Hz)*	*Time Accuracy (μs)*	*Measurement Transfer Time (ms)*	*Message Rate (Reports/sec)*	*Time Window (sec)*	*Data Loss Sensitivity (Reports or ms)*	*Performance Class (M/P/X/N)*	*Tools /Platforms*
Disturbance Analysis Compliance	0.01 p.u. mag 0.5 degrees 0.01 Hz	TVE	STD	0 - 30.0 Hz	STD	1000 ms	60 Reports/sec	Length of event	100 ms	M	
Frequency Response Analysis	0.5 degrees 0.01 Hz	TVE	STD	0 - 1.0 Hz	STD	1000 ms	5 Reports/sec	300 seconds	25 Reports	M	
Model Validation	0.01 p.u. mag 0.5 degrees 0.01 Hz	TVE	STD	0 - 30.0 Hz	STD	1000 ms	60 Reports/sec	Time frame of model	1000 ms	M	
Phasor Network performance monitoring & data quality	N/A	TVE	STD	0 - 30.0 Hz	STD	Measured	60 Reports/sec	86400 seconds	60 Reports	X	
Baseline Normal Phase Angle Trends	0.5 degrees	TVE	STD	0 - 10.0 Hz	STD	1000 ms	15 Reports/sec	86400 seconds	150 Reports	M	
Pattern Recognition/Correlation Analysis	0.01 p.u. mag 0.5 degrees 0.01 Hz	TVE	STD	0 - 0.2 Hz	STD	1000 ms	1 Report/sec	3600 seconds	20 Reports	M	
Situational Awareness Dashboard	0.01 p.u. mag 0.5 degrees 0.01 Hz	TVE	STD	0 - 1.0 Hz	STD	100 ms	30 Reports/sec	300 seconds	10 Reports	M	
Real Time Compliance Monitoring with Reliability Standards	0.01 p.u. mag 0.5 degrees 0.01 Hz	TVE	STD	0 - 1.0 Hz	STD	1000 ms	5 Reports/sec	1800 seconds	10 Reports	M	
Real Time Performance Monitoring and Trending	0.01 p.u. mag 0.5 degrees 0.01 Hz	TVE	STD	0 - 0.2 Hz	STD	1000 ms	1 Report/sec	3600 seconds	30 Reports	X	
Anomaly Characterization and Alarming	0.01 p.u. mag 0.5 degrees 0.01 Hz	TVE	STD	1.0 - 30.0 Hz	STD	100 ms	60 Reports/sec	3600 seconds	120 Reports	M	

Table 3-3: Requirements for Data Accuracy

Data accuracy element	Applications requirements
Measurement accuracy	Depending on the individual application, the requirement of the measurement accuracy must be such that the output of the application is usable (e.g., wide area visualization may not be as sensitive to data accuracy as automated real-time control).
Agreement of measurement algorithm	An application may have enough sensitivity to a particular measurement that it may require all PMUs to use the same measurement algorithm for a specified data channel (e.g., ROCOF).
Attribute accuracy	The correctness of other attributes attached to the measurement; in the case of PMUs, this primarily means the **temporal accuracy –** the agreement between the timestamp assigned to a measurement and the actual temporal coordinates at the instance of measurement. ● What happens to the application when there is a missing timestamp? Multiple missing timestamps? ● What happens when some of the timestamps have error? Will this cause an accumulating error in how the application treats phase? ● If the application is sensitive to errors of *temporal accuracy*, does it have a feature for handling measurements tagged by the PMU as having questionable timestamps (e.g., locally generated timestamps)? Attribute accuracy includes: ● **Geospatial accuracy** – the agreement between encoded and actual location coordinates of the PMU on the earth's surface. ● **Topological accuracy** – the agreement between the recorded location of the PMU and its actual location in the topology of the power system.
Data mislabeling	What will the application do if a data channel is mislabeled? Will it process it anyway and produce strange results? Will it report having done so? What will be the impact on the application's results? Will it skip over data judged to be out of bounds? Will it report having done so? What will be the impact on the application's results?

Table 3-4: Requirements for Data Content and Data Coverage

Data content/ coverage element	Locations in the *data process path* this can be impacted	Applications requirements
Application minimum geospatial span	● At the PMU ● At any aggregator	● What happens to the application if the *geospatial span* is not as expected? ● Is an application intended for a particular area (e.g., an area rich in renewables) and not relevant without a particular penetration of variable generation?
Application minimum topological span	● At the PMU ● At any aggregator	● What happens to the application if the *topological span* is not as expected? ● Is an application intended for a particular area (e.g., an area with particular features, such as very long lines, or a lack of very long lines)? ● Does the application require redundancy of *topological coverage*, such as for a state estimation algorithm? What happens if there is no redundancy of *topological coverage*?

Requirements an application may have for logical consistency are considered in Table 3-5. An application's underlying methods may tolerate gaps without difficulty, but start to lose accuracy in its results. Some *maximum gap rate* needs to be defined to give an idea of what is tolerable before the application's results become untrustworthy for a given period of the data. In addition, an application may have a largest gap in the data which it can tolerate before ceasing to function. For some applications, this may be zero – the application may not tolerate gaps without producing an error.

Table 3-5: Requirements for Logical Consistency

Logical consistency element	Locations in the data process path that could affect this	Applications requirements
Expected set of headers/identifiers	• At the PMU • When transmitted out via communications networks • At any aggregator • When read into a database • When loaded into the application	• What happens to the application if the headers and identifiers differ from what is expected? • What does the application do if headers or identifiers are missing? • Will an application correctly identify the situation (and error out or report a problem) if channels are mislabeled? Or will it treat those channels as correctly labeled, but with inaccurate measurements? • What will happen to the application if two channels are duplicated? Will it pick one? Will it try to use both? Will it cause errors?

3.3 Special Characteristics of the Application – "Training" the Application

Many types of analysis require that the application be "trained" using a minimum set of historical data, called the **application training period**, to enable the application to accurately process new data and the analyst to correctly understand and interpret the results. In many such cases, the analysis method becomes more accurate when that application training period is longer than the minimum historical data period recommended by the application creator. For example, a power systems analysis method that is trained on summer data will produce the most accurate results on more summer data, but may produce less accurate answers for other seasons until it has been trained on a full year or more of system data.

For a given combination of application and data, the analyst should consider how long an application training period[17] is appropriate. This can be complex and require some judgment calls to be made; for example, if the power system assets and topology have changed significantly over time, it may be that only results (and therefore, system operations data) from the past two or three years should be used for training the application. Alternatively, there may have been a change in data gathering or data storage practices that make the data incompatible before a certain historical point, or there may have been too few PMUs and clean data networks available before a certain point in time, making earlier PMU data of limited value for use for a specific application or analysis.

The following questions need to be answered for each application:

- Does the application require a minimum application training period or historical data span for application training? If so, how long is that period?

What I want	What I have
minimum application training period	available application training period
PMU minimum application training period	PMU available application training period
aggregator minimum application training period	aggregator available application training period

[17] Formally called an "available analytic history" in a previous version of this document.

- Does the application require that the application training period be recent? What happens if there is a significant break? For example, can a method that will start producing real-time results in 2016 be adequately trained on 2014 data? What about 2011 data?

What I want	What I have
minimum application training period length	available application training period length
PMU minimum application training period length	PMU available application training period length
aggregator minimum application training period length	aggregator available application training period length

- What is the ideal data span for an application training period? What is the recommended data span, which may be less than the ideal, but is more likely to be feasible?

What I want	What I have
maximum application training period age	available application training period age
PMU maximum application training period age	PMU available application training period age
aggregator maximum application training period age	aggregator available application training period age

- What happens to the results when the data available for an application training period is less than the ideal in time span, geospatial or topological scope, or quality?

For an application to be most accurate in a given situation, it will have been trained using the most recent available data covering a period long enough to contain numerous events of relevance to the analytical goals of the application, with topological consistency to the current and anticipated power system. It is valuable to know whether the application of interest has received sufficient “training” with an application training period that exceeds the recommended amounts and type of data. It is also important to know whether the current data and grid conditions to be analyzed fall within the scope of the data used for the application training period. As noted above, if the current data and conditions fall too far outside the scope that the application understands, the application may fail to produce trustworthy answers.

3.4 Requirements for the Data Source and Transformations Applied to the Data

In support of other requirements an application may have, the user may determine that the application has specific requirements for the data source. For example, in order to guarantee a certain confidence in accuracy, a user may specify a requirement that a PMU be of a particular PMU type, adhere to a particular standard, or even be of a specific model and firmware revision.

A particularly subtle and important issue that may impact application requirements is that of transformations applied to the data, including but not limited to compression algorithms.

4.0 The Impacts of Data Issues on Specific Applications

The PMU Applications Requirements Task Force (PARTF) recommends testing application vulnerabilities to data-quality problems by performing repeated tests of an application using a wide variety of flawed data inputs. This approach will help identify an application’s vulnerabilities to different types of data problems and help find the envelope of performance results relative to an application’s performance using perfect data. The proposed testing approach is to construct sets of PMU data that have been designed specifically to contain the various types of errors described in this document. The application would be tested using each of these input datasets, tracking the application’s results for each set of data flaws. Once the initial testing is completed, the analyst would look at the application’s analytical results to determine the impacts of different types, magnitudes, and combinations of data flaws. This would identify the types of data flaws that create the greatest analytical vulnerabilities for the application, and how sensitive the application’s performance is to those flaws. The parameters of the performance envelope (i.e., of application results relative to data errors) will be determined by where the application being tested still provides sufficiently accurate output (measured in terms of predefined analytical performance thresholds). The walls of this envelope will be linear inequalities separating acceptable errors and cumulative performance from unacceptable ones.

The envelope of application output results reflects the influence of multiple overlapping errors on the application performance. This idea is based on multi-dimensional PMU error analysis. By analyzing the application's performance envelope with multiple datasets, it will become clear how PMU data quality affects application performance and, thus, when data quality is insufficient to yield trustworthy application results.

4.1 Creating the Application Test

To construct an application's test, it is best to begin with either pseudo-PMU signals (obtained from dynamic simulations) or data from PMUs measured under controlled laboratory characteristics. These data sources should have well-understood characteristics (e.g., zero-noise signals), and should be free of unknown or uncontrolled measurement error. The set of PMU signals will be modified to add certain levels of errors (e.g., certain levels of noise or frequency error). Different types and combinations of data errors, quality problems, and losses will be tested sequentially, and each compared to the original (perfect) dataset to determine how each set of data flaws affects the application's performance.[18]

This process will allow us to determine the impacts of PMU and data process path flaws on the performance of specific synchrophasor applications. Note that this simulation must reflect the entire synchrophasor data process path, not just the impact of errors generated inside the PMU. Figure 4-1 shows a flowchart for testing synchrophasor applications.

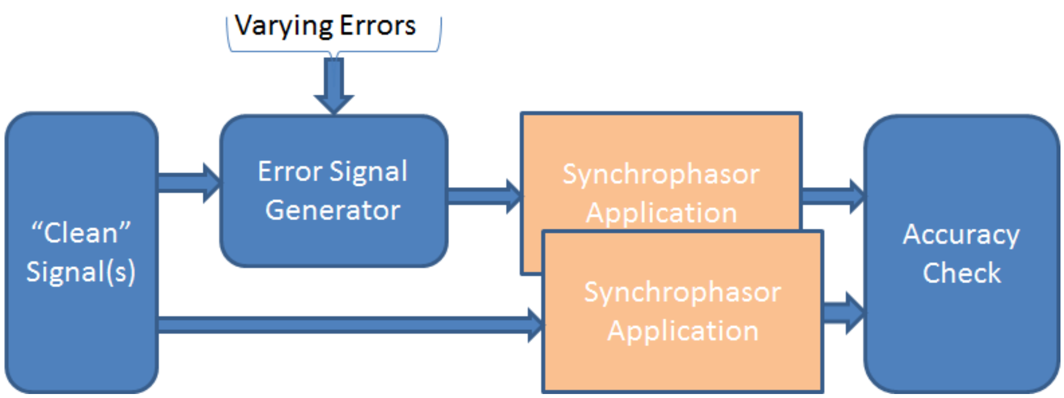


Figure 4-1: Methodology Flowchart

The process begins with a "clean" set of data, produced from PMUs or by dynamic simulations of power system transients. The clean dataset is assumed to represent the closest PMU data representation to the unaltered physical signal and to lack any significant errors or weaknesses that would affect the application's performance analysis. This means that our approach is based on a comparison of application's results with added error against the (assumedly) clean data. This set of clean PMU data is sent to the tested application, and the result (i.e., the damping ratio or frequency response) is stored for a

[18] Work conducted by the National Institute of Standards and Technology (NIST) and the Pacific Northwest National Laboratory (PNNL) (Huang2015) suggested that real PMUs do not display simple Gaussian error patterns. Rather, PMU measurement errors are quite systematic and are related to frequency and ROCOF. Some research has been considering the design of a PMU model that can have parametrically controlled error response. This can actually be operated with simulated and actual analog input. NIST already has a model of the C37.118.1 Annex C model that can run in a simulation and with analog input.

subsequent accuracy check. The same clean data is then distorted repeatedly using varying data errors and omissions that could affect the application's performance. Each distorted dataset is fed to the application in its turn, and the analytical result captured. Then the difference between all of the error-based results is compared against application-specific performance thresholds. The evaluator can use this comparison to decide whether the analytical result from each distorted dataset is acceptable or not, to build the data-quality performance envelope for the application.[19]

The original "clean" dataset of measurements from multiple PMUs would be modified to test the impacts of numerous types of data flaws and errors upon the tested application's performance. Some of those modifications might include the following:

- different types of measurement errors
- inserting data losses and drop-outs to test the impact of data sparsity and the loss of specific PMUs
- mislabeling data channels
- erroneous or duplicative timestamps
- creating delays in data delivery to the application.

Figure 4-2 represents a hypothetical example of the application testing results. The application is tested against three types of errors – *Err1*, *Err2,* and *Err3*. The green points and red points are test samples; green points represent acceptable performance and red points represent unacceptable performance. The blue boundary line separates green points from red points and indicates the performance envelope of the tested application.

Development of an application's performance envelope may be more of an art than a science. The process should take many iterations and use many test datasets, as described above. Analysts should use knowledge of relevant systems and of the application to assess the impact and significance of different combinations of data flaws. Repeated tests should identify the types of data flaws that cause the greatest problems for effective application performance, which should inform application tuning and improvement. This process might also point to whether improvements in the data network or the application data preprocessor are needed to reduce the application's vulnerability to specific data flaws.

It is also worth noting that an additional probabilistic performance test methodology is needed for those applications that do not have a deterministic "right" answer from a repeatable process (e.g., a mode meter).

[19] It is up to the user to determine the acceptable bounds of accuracy and performance for a specific application or study. These may vary between deployment environments, revisions of software, and even the importance of the application in an operational or analysis scenario.

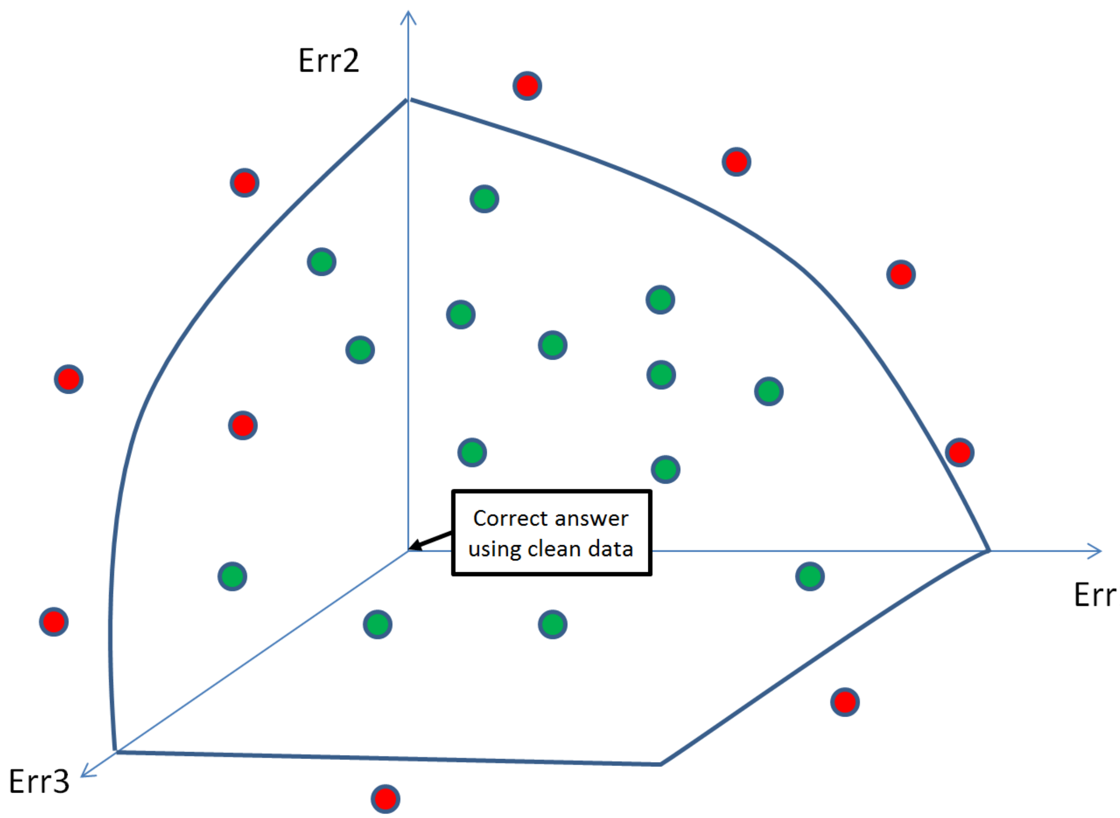


Figure 4-2: Application Performance Envelope for Varying Data Errors Relative to Clean Data

4.2 An Example of Examining the Impact of the Network on PMU Data Quality

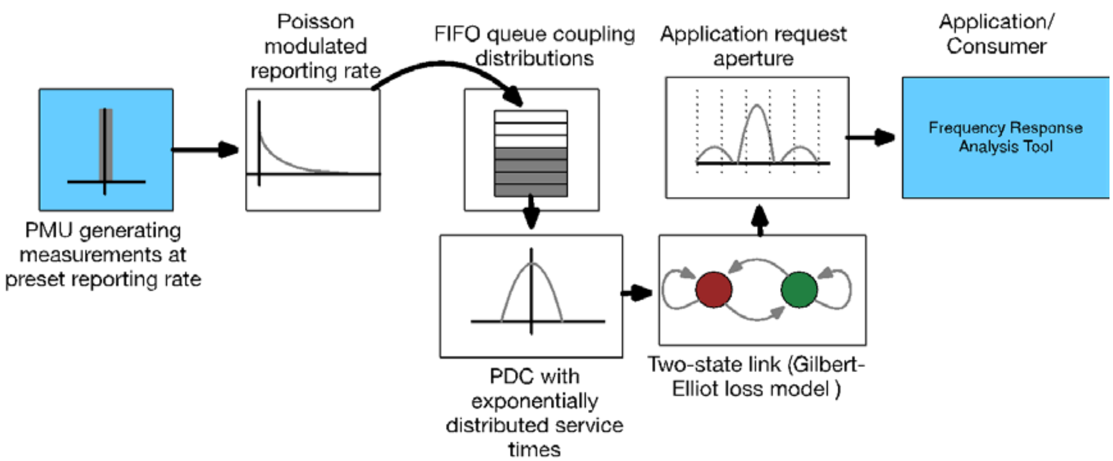
A prudent first step in analyzing network impacts on PMU data quality is to fully understand the constraints of the desired network (e.g., frame losses, data delivery delays, and mean and maximum gap times). This step may require translating the definitions from network parlance into the terms used in this document. Further, the network analysis method may need to be altered to follow the analysis method proposed in the text box below, to identify which network performance characteristics might compromise data sufficiently to compromise application performance.

Consider the impact of quality of service (QoS) elements on network performance. For example, for a network with QoS requirements in place, the end-to-end latency for Class A data might be specified as 50 ms, of which 25 ms is reserved for data transport through the WAN and 12.5 ms each for data processing at a publishing and subscribing PDC. Class A data is also assured to be delivered to an application with 99.9999% probability, with a maximum service interruption time of 5 ms. It would not be simple to map and test such a QoS specification in the parametric study required to generate an application's data quality performance envelope. In addition, impacts of additional data stream parameters (e.g., latency impacts of aggregator processing and network-caused data drop-outs) should be examined.

Example Study Approach for Network Influences

The figure below illustrates one possible approach to studying data process path effects using composite stochastic processes. The model uses a Poisson arrival process to model measurement rate modulation due to network access effects and delays in the data bus. A single data queue couples the Markovian arrival process (a mathematical model for the time between job arrivals to a system) with an exponential distribution of service times. Service time distribution captures the variation introduced by data aggregators, processors, PDCs, etc. This model format typically called an M/M/1 queue, as discussed in Kihl (2011). Finally, a two-state Markov approach is used to characterize burst-noise due to error detection and recovery mechanisms in transmission channels. Because each model parameter is explicitly represented in the final expression for the full model, this approach enables the conduct of parametric surveys to generate a spectrum of dataset modifications that could be used to identify a performance envelope for this set of network errors.

An abstract representation of the network focusing on network parameters that may be mapped to PMU data quality definitions



Significant changes in application requirements, network components, or in PMUs themselves may require repeating the end-to-end validation of the application and data process stream with use cases to ensure achievement of the performance goals. Applications using a specific network as a component in their architecture do not naturally inherit its performance assurance standards; instead they have to depend on extensive simulation-based testing of use cases to validate their performance. These analyses use commercial network simulation tools, which offer the user a large set of tunable parameters to simulate the network at great detail.

5.0 Conclusions and Next Steps

This document has offered definitions for a broad suite of data attributes and data quality problems, and proposed an overall methodology to evaluate the impact of weaknesses in PMU data on synchrophasor applications.

Clearly, a wide variety of data problems exist, and it is reasonable to expect that different data flaws affect application performance in varying ways. Further, different applications likely have differing sensitivities to data flaws (e.g., an application using an incoming data stream is certainly far more vulnerable to data problems than one that uses data off-line). And few applications can work effectively without the benefit of training with historical data to assure trustworthy results.

The authors look forward to working with members of the synchrophasor community to build agreement on common terminology and definitions for data quality. Using the process outlined in this paper to test an application using a suite of related, but error-modified, datasets could reveal new insights into application sensitivities and vulnerabilities. NIST staff are developing clean datasets and data modification methods to be used in the future to implement and test synchrophasor applications' data accuracy and availability. The PARTF team hopes that this focus on data quality definitions and test methods will lead to a better appreciation for the complexity and impacts of synchrophasor data weaknesses and error on the end-use application.

Bibliography

Liu 2015 Yilu Liu et al., PMU Error Impact on Measurement-Based Applications, CERTS, June 10, 2015, Washington, DC, http://www.energy.gov/sites/prod/files/2015/07/f24/7.%20Liu%20Synchrophasor-based%20applications.pdf

Meliopoulo 2007 A. P. Meliopoulos, et al., Synchrophasor Measurement Accuracy Characterization, NASPI Performance & Standards Task Team, 2007, pp. 43-58.

Huang 2008 Z. Huang, B. Kasztenny,V. Madani, Fellow, K. Martin, Fellow, S. Meliopoulos, D. Novosel, and J. Stenbakken, “Performance Evaluation of Phasor Measurement Systems,” IEEE Power Engineering Society General Meeting 2008, Pittsburgh, PA.

Kirkham 2015 Kirkham, H., *A Conceptual Framework for Measurement*, Technical Report, *PNNL-24071*, Richland, WA: Pacific Northwest National Laboratory, 2015.

Hu 2010 Hu, Y., et al. "NASPInet specification-an important step toward its implementation", *2010 43rd Hawaii International Conference on System Sciences.* IEEE, 2010.

Hasan 2009 Hasan, Ragib, Rakesh Bobba, and Himanshu Khurana. "Analyzing NASPInet data flows." *Power Systems Conference and Exposition, 2009. PSCE'09. IEEE/PES*. IEEE, 2009.

Kihl 2011 Kihl, Maria, et al. "Performance measurements and modeling of database servers." *Sixth International Workshop on Feedback Control Implementation and Design in Computing Systems and Networks (FeBID 2011)*. 2011.

Cavin 2011 Cavin, David, Yoav Sasson, and André Schiper. "On the accuracy of MANET simulators." *Proceedings of the second ACM international workshop on Principles of mobile computing*. ACM, 2002

USGS 1998 United States Geological Survey, “Spatial Data Transfer Standard”, ANSI NCITS 320-1998, 1998, http://mcmcweb.er.usgs.gov/sdts/.

Guptill 1995 Elements of Spatial Data Quality. Stephen C. Guptill, JOEL L AUTOR MORRISON. Elsevier Science Limited, 1995 – Science – 202 pages.

Goldstein 2015 NISTIR 8077: Investigation of PMU Response to Leap Second: 2015 Investigation of PMU Response to Leap Second: 2015. Allen Goldstein, Dhananjay Anand, Ya-Shian Li-Baboud. National Institute of Standards and Technology, 2015. http://dx.doi.org/10.6028/NIST.IR.8077

Amidan 2014 *An Initial Application of DISAT (Data Integrity Situational Awareness Tool*). BG Amidan, JD Follum, KA Freeman, JE Dagle. PNNL, December 2014.

Buckley 1997 Buckley, David J. *The GIS Primer: An Introduction to Geographic Information Systems*. 1997 http://bgis.sanbi.org/GIS-primer/index.htm

Huang 2015 H. Huang, N. Zhou, M. Anitescu, S. Wang, “Exploring state estimation techniques to accommodate non_Gaussian noise,” in Proc. SIAM Conference on Computational Science and Engineering, Salt Lake City, UT, March 14-18, 2015 https://www.pathlms.com/siam/courses/1043/sections/1281/documents/1144

NASA 1994 Metrology — Calibration and Measurement Processes Guidelines. NASA Metrology and Calibration Working Group. NASA Reference Publication 1342, 1994. http://www.hq.nasa.gov/office/codeq/doctree/NASA_RP1342.pdf

Modell 1994 Martin E. Modell, Data Analysis, Data Modeling, and Classification, McGraw Hill Software Engineering Series, 1992. https://books.google.com/books/about/Data_Analysis_Data_Modeling_and_Classifi.html?id=wthQAAAAMAAJ

PJM 2014 PJM, "Attachment M-TO Registry File Template", Operating Committee Meeting 04-08-2014 notes. [Online]. Available: http://www.pjm.com/~/media/committees-groups/committees/oc/20140408/20140408-item-06-attachment-m-to-registry-file-template.ashx.

Appendix A – Tables of Data Quality Terms, Metrics, and Application Requirements

Attributes of single data points / measurement specifiers (*The minimum attributes necessary to specify a measurement)*			**Attribute affected by**		
Name of single data attribute	**Metric / specification**	**Related application requirements**	**The PMU?**	**Communications networks?**	**Aggregators?**
Signal being measured	A quantity the measurement will represent, specified by an agreed-upon definition, usually from a standard.		Y	N	N
Measured result	The numerical value of the measurement, expressed as a number.		Y	N	N
Standard units	The units of measure for the *measured result*, (e.g., current measured in amperes), expressed by the standard abbreviations for the units in question.	An application may require that a given measurement or data channel use specified *standard units.*	Y	N	N
Precision	A specified range within which the *measured result* falls (e.g., to three decimal points), indicating the repeatability of the measurement and measurement process, expressed as a number of decimal points. Note that *precision,* which quantifies the repeatability of a particular measurement/measurement process, is independent of *accuracy,* which quantifies how close a *measured result* is to the *signal being measured.* A measurement process can be accurate but not precise, if its *measured results* are centered around the *signal being measured,* but not closely grouped. A measurement process could also be precise but not accurate, if it returns closely matching *measured results* but those *measured results* are all substantially incorrect representations of the grid condition.	An application may require that a given *measured result* be given to a specified *precision.*	Y	N	N
Confidence interval	A specified fraction or percentage (e.g., the calculated measurement comes within +/- 2% of the signal being measured) which indicates the certainty with which the *measured result* falls within the stated *precision,* usually expressed as a percentage (+/- X%).	An application may require that a given *measured result* be specified to a specified *confidence interval.*	Y	N	N

Attributes of single data points / measurement accuracy / source influence *(Source influence is created by phenomena within the power system that could cause the measured result to differ from the signal being measured, e.g., harmonic interference in the power system)*			**Attribute affected by**		
Name of single data attribute	**Metric / specification**	**Related application requirements**	**The PMU?**	**Communications networks?**	**Aggregators?**
Noise in the power system	Noise in the source signal in terms of wideband SNR, expressed as a percentage or in decibels.	An application may require that noise in a particular signal be below a certain percentage or number of decibels.	Y	N	N
Out-of-band interfering signals	Interfering signals present in the source signal in terms of narrowband SNR, expressed as a percentage of nominal magnitude.	An application may require that *interfering signals* not exceed a certain percentage of nominal magnitude.	Y	N	N
Harmonics	Total harmonic distortion (*THD*) of the source signal, expressed as a percentage of nominal magnitude.	An application may require that *THD* not exceed a certain percentage of nominal magnitude.	Y	N	N
Instrument channel errors	Hardware affecting PMU measurements: - Potential transformers (PTs) - Current transformers (CTs) - Capacitively coupled voltage transformers (CCVTs) - Electro-optical voltage transformers (EOVTs) - Magneto-optical current transformers (MOCTs) - Cables - Differences between measurement-class versus protection-class instrument transformers	An application may require that supportive hardware allow the PMU to produce results within a certain variance from a reference signal.	Y	N	N
Global Positioning System (GPS) loss, spoofing, or meaconing	How this is expressed varies. GPS signals can be lost entirely, but they can also be interfered with to produce a false input signal. This is a type of *attribute accuracy*, specifically *temporal accuracy.*	An application may require that GPS timestamps be accurate to within a specified band.	Y	N	N

Attributes of single data points / measurement accuracy / induced error (*Induced error is error that is introduced by the PMU estimation process*)			**Attribute affected by**		
Name of single data attribute	**Metric / specification**	**Related application requirements**	**The PMU?**	**Communications networks?**	**Aggregators?**
Synchrophasor total vector error (TVE)	*Synchrophasor TVE* is the magnitude of the synchrophasor's RMS amplitude/phase error vector, expressed as a percentage, as per IEEE Std. C37.118.1-2011, C37.118.1a-2014.	An application may require that *synchrophasor TVE* remain within a specified percentage.	Y	N	N
Synchrophasor magnitude error	The difference between the RMS amplitude of the synchrophasor estimate and the RMS amplitude of the PMU's input signal, expressed as a percentage.	An application may require that *synchrophasor magnitude error* remain within a specified percentage.	Y	N	N
Synchrophasor phase error	The difference between the reported synchrophasor angle and the absolute angle of the PMU's input signal. *Synchrophasor phase error* can be expressed as a percentage, or specified in degrees or radians.	An application may require that *synchrophasor phase error* remain within a specified percentage or within a specified number of degrees or radians.	Y	N	N
Frequency error (FE)	*FE* is the difference between the reported frequency and the instantaneous frequency of the PMU's input signal, expressed in Hertz (Hz).	An application may require that *FE* remain within a specified fraction of a Hz.	Y	N	N
Rate of change of frequency (ROCOF) error	*ROCOF error* is the difference between the reported ROCOF and the ROCOF of the PMU's input signal, expressed in Hz/second.	An application may require that *ROCOF error* remain within a specified fraction of a Hz/s.	Y	N	N

Attributes of single data points / attribute accuracy			Attribute affected by		
Name of single data attribute	**Metric / specification**	**Related application requirements**	**The PMU?**	**Communications networks?**	**Aggregators?**
Temporal accuracy	Temporal accuracy refers to the agreement between the timestamp assigned to a measurement and the actual temporal coordinates at the instant of measurement. Note that a PMU or aggregator may have a process to substitute a locally-generated timestamp in case of failure of the GPS signal at the PMU or if a timestamp is missing at the aggregator. GPS spoofing or a satellite timing error could cause a failure or error in the recorded temporal accuracy of a PMU measurement.	An application may require that timestamps be accurate to within a specified band. This could mean that any substitute timestamp generation process employed in the event of the loss of a GPU signal would need to be accurate to the specified application requirement.	Y	N	N
Geospatial accuracy	*Geospatial accuracy* refers to the agreement between encoded and actual location coordinates of the PMU on the earth's surface.	An application may require that that *geospatial accuracy* of a PMU's location be within a specified band.	Y	N	N
Topological accuracy	*Topological accuracy* refers to the agreement between the recorded location of the PMU and its actual location in the topology of the power system.	An application may require that that *topological accuracy* of a PMU's location be within a specified band.	Y	N	N

Attributes of single data points / data lineage (i.e., metadata) / data source*			Attribute affected by		
Name of single data attribute	**Metric / specification**	**Related application requirements**	**The PMU?**	**Communications networks?**	**Aggregators?**
PMU type	"*Protection class*" or "*measurement class*" specified by the manufacturer for each PMU that contributes to a dataset.*	An application may require a specific *PMU type*. An application's requirements for other data attributes may disallow a specified *PMU type.*	Y	N	N
PMU standard followed	Version of IEEE C37.118.1 or other standard complied with by each PMU that contributes to a dataset.*	An application may require adherence to a particular standard or version of that standard.	Y	N	N
PMU model, firmware revision, and configuration settings	Model name and part number, and firmware revision specified by the manufacturer for each PMU that contributes to a dataset; a list of user-set configuration settings for each PMU that contributes to a dataset. Keeping this information recorded in a manner associated with a dataset contributes to troubleshooting when unexpected problems arise. Some details may seem inconsequential for any foreseeable use when they are selected, but may later prove crucial to understanding results of a more recently added type of application or analysis.*	Application requirements for a number of metrics may disallow data from certain combinations of PMU models, firmware revisions, and configuration settings.	Y	N	N
PMU-supplied data headers	Headers specified by the manufacturer and/or specified by the programmer for each PMU that contributes to a dataset.*	An application may require a measurement be delivered with certain headers.	Y	N	N
Aggregator-supplied data headers	Headers specified by the manufacturer and/or specified by the programmer for each aggregator that contributes to a dataset. Note that a header or set of headers can be lost at any aggregator or communications network through which the data passes, and that an aggregator may output a header set that differs from one or more of the PMUs from which it takes input.*	An application may require a measurement be delivered with certain headers.	N	N	Y

* Assess regularly – this attribute does not need to accompany every measurement, but should be weighed when determining the suitability of data for an application or study or when troubleshooting problems with application results.

Attributes of single data points / data lineage (i.e., metadata) / data coverage*			Attribute affected by		
Name of single data attribute	**Metric / specification**	**Related application requirements**	**The PMU?**	**Communications networks?**	**Aggregators?**
PMU geospatial location	Geospatial coordinates for each PMU that contributes to a dataset, specified in degrees:minutes:seconds north or south and degrees:minutes:seconds east or west.*	An application may require data from a particular *PMU geospatial location.*	Y	N	N
PMU topological location	Grid topological location from a specific PMU, specified in some manner that is meaningful to the service area in question.*	An application may require data from a particular *PMU topological location* in the power grid.	Y	N	N
Attributes of single data points / data lineage (i.e., metadata) / data transformation methods (*Any interpolation, downsampling, correction, compression and/or any other type of mathematical transformation***)***					
Transformations applied to the data at the PMU	For each PMU that contributes to a dataset, whether or not any interpolation, downsampling, correction, and/or any other type of mathematical transformation was applied to any of the data, and specifically 1) how, and 2) which data.*	An application may require that the PMU not make any such interpolation, downsampling, correction, and/or any other type of mathematical transformation. An application may require that some data modifications are allowable but others are not.	Y	N	N

* Assess regularly – this attribute does not need to accompany every measurement, but should be weighed when determining the suitability of data for an application or study or when troubleshooting problems with application results.

Attributes of datasets / data lineage (i.e., metadata) / data transformation methods (*Any interpolation, downsampling, correction, compression and/or any other type of mathematical transformation*)*			**Attribute affected by**		
Name of single data attribute	**Metric / specification**	**Related application requirements**	**The PMU?**	**Communications networks?**	**Aggregators?**
Transformations applied to the data at an aggregator	For each aggregator that contributes to a dataset, whether and how any interpolation, downsampling, correction, compression and/or any other type of mathematical transformation was applied to any of the data in a dataset, and specifically 1) how, 2) which data, 3) where in the *data process path* this transpired. This includes interpolation or downsampling to provide a single *aggregator reporting rate* when contributing PMUs have different *PMU reporting rates.**	An application may require specific changes to data not be made after the data is generated at the PMU. An application may require that data not have been compressed, not have been compressed beyond a certain point, or that specific compression algorithms not be used.	N	N	Y
Transformations applied to the data during archiving	It is important to understand and track any transformations or changes to data made during archiving, in addition to *transformation methods applied to the data at the PMU* and *transformation methods applied to the data at an aggregator.* It is not uncommon for power system data to be archived in multiple stages. For example, data less than three months old might reside in one database, then at three months be moved to another archive, then at a year moved to yet another archive for longer-term storage. Each such re-archiving of the data may involve moving it to a new system and applying different types of data compression. As a result it is possible to retrieve what should be identical data from different archives and get completely different results using that data. Additionally, an analysis technique may cease to function on older data simply because the older data has suffered more through compression.*	An application may require specific changes to data not be made after the data is generated at the PMU. An application may require that data not have been compressed, not compressed beyond a certain point, or that specific compression algorithms not be used.	N	N	Y

* Assess regularly – this attribute does not need to accompany every measurement, but should be weighed when determining the suitability of data for an application or study or when troubleshooting problems with application results.

Attributes of datasets / data lineage (i.e., metadata) / data coverage*			**Attribute affected by**		
Name of single data attribute	**Metric / specification**	**Related application requirements**	**The PMU?**	**Communications networks?**	**Aggregators?**
Aggregator geospatial coverage	Total geospatial coverage for each aggregator that contributes to a dataset. Specified in degrees:minutes:seconds north or south and degrees:minutes:seconds east or west. The effective *aggregator geospatial coverage* can be impacted at the PMU if one or more PMUs cease reporting, so that the *geospatial coverage* of the dataset is lowered.*	An application may require an **application minimum geospatial span** in order to deliver meaningful results. A study or application may require *aggregator geospatial coverage* of a particular area.	Y	N	Y
Aggregator topological coverage	Total grid topological coverage for each aggregator that contributes to a dataset, specified in some manner that is meaningful to the service area in question. The effective *aggregator topological coverage* can be impacted at the PMU if one or more PMUs cease reporting, so that the topological coverage of the set is lowered.*	An application may require an **application minimum topological span** in order to deliver meaningful results; how that span is defined needs to make sense for both the application itself and for the features of each service area the application will analyze. An application may require redundancy of some or all *aggregator topological coverage;* analytical methods such as state estimation or local state estimation are examples of applications for which some redundancy of *aggregator topological coverage* may be required.	Y	N	Y

* Assess regularly – this attribute does not need to accompany every measurement, but should be weighed when determining the suitability of data for an application or study or when troubleshooting problems with application results.

Attributes of datasets / data lineage (i.e., metadata) / data coverage*			**Attribute affected by**		
Name of single data attribute	**Metric / specification**	**Related application requirements**	**The PMU?**	**Communications networks?**	**Aggregators?**
PMU temporal coverage	• **PMU temporal coverage** is the period(s) of time for which data is available from a specific PMU – the time period(s) for which the PMU is in operation. • **Aggregator temporal coverage** is the period(s) of time for which data is available from a specific aggregator – the time period(s) for which that aggregator is in operation. Both of these attributes still allow for the possibility of gaps – see *gap rate*, *gap size*, and *largest known gap).* Some part of the *PMU temporal coverage* or *aggregator temporal coverage* can seem effectively lost at any aggregator or communications network through which the data passes, but note that it may be possible to retrieve missing data from stages in the *data process path.* Note that a newer PMU may not have the *PMU temporal coverage* of other PMUs in a dataset.*	An application may use an underlying mathematical method or set of mathematical methods that require an **application training period** of historical data to train (and possibly test) the method(s). (If both training and testing data are required, those must be two distinct and non-overlapping periods of data, but for simplicity, the whole is referred to as the *application training period.)* The *application training period* is distinct from the *application time window* an application may require of an incoming PMU data stream. • The **PMU minimum application training period** is the minimum amount of historical *PMU temporal coverage* that must be available prior to the actual period to be analyzed. • The **aggregator minimum application training period** is the minimum amount of historical *aggregator temporal coverage* that must be available prior to the actual period to be analyzed.	Y	Y	Y
Aggregator temporal coverage			Y	Y	Y

* Assess regularly – this attribute does not need to accompany every measurement, but should be weighed when determining the suitability of data for an application or study or when troubleshooting problems with application results.

Attributes of datasets / data lineage (i.e., metadata) / data content*			Attribute affected by		
Name of single data attribute	**Metric / specification**	**Related application requirements**	**The PMU?**	**Communications networks?**	**Aggregators?**
Number of PMUs	Total number of PMUs in a dataset or total number of PMUs providing input to an aggregator.*	An application may require an **application minimum number of PMUs** in order to produce meaningful results. From the point of view of the application if one or more PMUs cease reporting, the effective *number of PMUs* goes down. As a result, the *number of PMUs* can be impacted at the PMU if one or more PMUs cease reporting unless it is possible to retrieve missing data from stages in the *data process path.*	Y	Y	Y
Channels provided by the PMU	List of which data channels (signals being measured) are provided for each PMU that contributes to a dataset.*	An application will require certain data channels (signals being measured) to perform its analysis. Note that from the point of view of the application, the *channels supplied by the PMU* or the *channels supplied by the aggregator* can appear to change if there is a PMU failure, an aggregator failure, or a failure in the communications network, unless it is possible to retrieve missing data from stages in the *data process path.*	Y	Y	Y
Channels provided by the aggregator	List of which data channels (signals being measured) are provided for each aggregator that contributes to a dataset, including aggregator metrics that are included in the data streams. A single PMU produces a specified set of data channels. A PMU dataset, such as those PMUs attached to a single aggregator or all PMUs for a service territory, may have a slightly differing subset of channels for which data is available.*		N	N	Y

* Assess regularly – this attribute does not need to accompany every measurement, but should be weighed when determining the suitability of data for an application or study or when troubleshooting problems with application results.

Attributes of datasets / data lineage (i.e., metadata) / data content*			Attribute affected by		
Name of single data attribute	**Metric / specification**	**Related application requirements**	**The PMU?**	**Communications networks?**	**Aggregators?**
PMU reporting rate	Measurements per second specified by the manufacturer and/or specified by the programmer for each PMU that contributes to a dataset, recorded in samples per second, e.g., 30 samples/second or 60/samples/second. The *PMU reporting rate* is assumed to be constant (while a variable rate is possible, this is not done in practice). Note that the *PMU reporting rate* is distinct from the *message rate* of an incoming PMU data stream; the *PMU reporting rate* is how often a given PMU is set to take a measurement, while the *message rate* of an incoming PMU data stream has to do with how often the stream is expected to deliver a new message – these two rates may be the same, but realities of measurement and network transport mean that they can and do differ, particularly when there is a problem.*	An application may require an **application minimum reporting rate** to be effective. An application may have an **application maximum reporting rate** that it can effectively absorb. From the point of view of the application, the effective *PMU reporting rate* can be changed if the PMU suffers gaps beyond the tolerance of the application, unless it is possible to retrieve missing data from stages in the *data process path.*	Y	N	N
Aggregator reporting rate	Measurements per second specified by the manufacturer and/or specified by the programmer for each aggregator that contributes to a dataset, recorded in samples per second, e.g., 30 samples/second or 60/samples/second. The aggregator reporting rate is assumed to be constant (while a variable rate is possible, this is not done in practice). Note that the aggregator reporting rate output by an aggregator might differ from the PMU reporting rate of one or more of the PMUs for which it takes input.*	An application may require an **application minimum reporting rate** to be effective. An application may have an **application maximum reporting rate** that it can effectively absorb. From the point of view of the application, the effective *aggregator reporting rate* can be changed if the aggregator suffers gaps beyond the tolerance of the application, unless it is possible to retrieve missing data from stages in the *data process path.*	N	N	Y

* Assess regularly – this attribute does not need to accompany every measurement, but should be weighed when determining the suitability of data for an application or study or when troubleshooting problems with application results.

Attributes of datasets / logical consistency / metrology persistence and consistency (*Whether the metrology (or measurement method) remains the same over time (persistence) and across PMUs (consistency))**			**Attribute affected by**		
Name of single data attribute	**Metric / specification**	**Related application requirements**	**The PMU?**	**Communications networks?**	**Aggregators?**
PMU metrology persistence	Y/N: Whether or not the PMU's measurement techniques for each type of data gathered remain the same from timestamp to timestamp.*	An application may require *PMU metrology persistence* for particular data channels or for all data channels.	Y	N	N
Aggregator metrology persistence	Y/N: Whether or not the measurement techniques for each type of data gathered into an aggregator remain the same from timestamp to timestamp.*	An application may require *aggregator metrology persistence* for particular data channels or for all data channels.	Y	N	N
Aggregator metrology consistency	Y/N: Whether or not the measurement techniques for each type of data gathered remain consistent across all PMUs providing input to the aggregator.*	An application may require *aggregator metrology consistency* across all data channels.	Y	N	Y
Attributes of datasets / logical consistency / header persistence and consistency (*Whether the headers remain the same over time (persistence) and across PMUs (consistency). Unexpected headers, such as from older data predating conventions, could produce some highly unexpected results if not corrected.)**					
PMU header persistence	Y/N: Whether or not an individual PMU produces an identical set of headers for every data frame (all the information produced per timestamp).*	An application may require a consistent set of headers. Note that data for any given PMU may extend back before currently accepted practices on header specifications.	Y	N	N
Aggregator header persistence	Y/N: Whether or not an aggregator outputs an identical set of headers for every timestamp of its output*		Y	N	Y
Aggregator header consistency	Y/N: Whether or not an aggregator outputs an identical set of headers across all of its inputs. The *aggregator header consistency* can be impacted by communications errors.*		Y	Y	Y
Attributes of Datasets / Logical Consistency / Data Frame Persistence and Consistency A104 (*Whether the structure of data frames (all the information produced per timestamp) remain the same over time (persistence) and across PMUs (consistency))**					
PMU data frame persistence	Y/N: Whether or not an individual PMU produces data frames which contain the complete set of expected data and attributes for each timestamp.*	An application may require a complete frame with all measurements present, while other applications may only require particular data channels.	Y	N	N
PMU data frame persistence	Y/N: Whether or not an aggregator outputs data frames which contain the complete set of expected data and attributes for each timestamp.*		Y	Y	Y

* Assess regularly – this attribute does not need to accompany every measurement, but should be weighed when determining the suitability of data for an application or study or when troubleshooting problems with application results.

Attributes of datasets / logical consistency / metrology persistence and consistency (*Whether the metrology (or measurement method) remains the same over time (persistence) and across PMUs (consistency)*)*			**Attribute affected by**		
Name of single data attribute	**Metric / specification**	**Related application requirements**	**The PMU?**	**Communications networks?**	**Aggregators?**
Aggregator data frame consistency	Y/N: Whether or not an aggregator outputs the same set of expected data and attributes for each of its inputs. Note that the *aggregator data frame consistency* can be impacted by communications errors.*		Y	Y	Y
Attributes of datasets / logical consistency / data frame order consistency (*Whether the data frames at an aggregator are recorded in the correct order)***					
Aggregator data frame order consistency	Y/N: Whether or not the aggregator has data frames recorded in the correct order. Note that the aggregator data frame order consistency can be impacted by communications errors.*	An application may require that data frames are recorded in the correct order when being read into the application. (Note this may require a front-end preprocessor to check and rectify frame order.)	Y	Y	Y

* Assess regularly – this attribute does not need to accompany every measurement, but should be weighed when determining the suitability of data for an application or study or when troubleshooting problems with application results.

Attributes of datasets / logical consistency / standards compliance persistence and consistency (*Whether compliance to standards remains the same over time (persistence) and across PMUs (consistency)*)*			**Attribute affected by**		
Name of single data attribute	**Metric / specification**	**Related application requirements**	**The PMU?**	**Communications networks?**	**Aggregators?**
PMU standards compliance persistence	Y/N: Whether or not a PMU remains persistent over time as to which standards are adhered to.*	An application may require adherence to a particular standard or version of that standard, or that data be preprocessed through mapping software that maps data gathered under one standard to match data gathered under another.	Y	N	Y
Aggregator standards compliance persistence	Y/N: Whether or not the PMUs contributing input to an aggregator remain persistent over time as to which standards are adhered to.*		Y	N	Y
Aggregator standards compliance consistency	Y/N: Whether or not the PMUs contributing input to an aggregator are consistent among themselves as to which standards are adhered to.*		Y	N	Y
Attributes of datasets / logical consistency / reporting rate persistence and consistency A121 (*Whether the reporting rates remain the same over time (persistence) and across PMUs (consistency). These are attributes of the intended* reporting rate *for parts or whole of a dataset, without respect to gaps or drop-outs, which are covered in later sections.*)*					
PMU reporting rate persistence	=B158Y/N: B158Y/N: B158YDoes the data produced from the PMU continue to have the same *PMU reporting rate* over time, that is, does it continue to produce a measurement at every specified time interval? The *PMU reporting rate persistence* is a function only of whether the individual PMU continues to record a measurement at the specified time intervals. It does not include situations where the PMU recorded the measurement, but the expected method of gathering the measurement failed.*	An application may require that the *PMU reporting rate* not change over time.	Y	N	N
Aggregator reporting rate persistence	Y/N: Does the data output from an aggregator continue to have the same *aggregator reporting rate* over time?*	An application may require that the *aggregator reporting rate* not change over time.	N	Y	Y
Aggregator reporting rate consistency	Y/N: Does the set of PMUs providing input to an aggregator have a consistent *PMU reporting rate* across all PMUs in the set?*	An application may require that the set of PMUs providing input have a consistent *PMU reporting rate.*	N	Y	Y

* Assess regularly – this attribute does not need to accompany every measurement, but should be weighed when determining the suitability of data for an application or study or when troubleshooting problems with application results.

Attributes of datasets / data completeness / gap rate (*A gap is defined as the time from when measurements stop until measurements start again. A gap refers to data that is missing completely, not delayed in delivery due to communication drop-outs.*)**†***			**Attribute affected by**		
Name of single data attribute	**Metric / specification**	**Related application requirements**	**The PMU?**	**Communications networks?**	**Aggregators?**
Gap rate in data from PMU	The number of missing measurements or timestamps in the PMU data per unit time, expressed in gaps/second, gaps/day, or gaps/year. The *gap rate in data from PMU* can be nonzero at the PMU, increased at any aggregator, and increased by any communications network through which it travels.†*	An application's underlying methods may tolerate *gaps* without difficulty, but start to lose accuracy in its results. Some **application maximum gap rate** needs to be defined to give an idea of what is tolerable before the application's results become untrustworthy for a given period of the data. Note that the *application maximum gap rate* will apply to the gap rate of data after it has passed through the entire data process path before the application, not just to the *gap rate in data from PMU.*	Y	Y	Y
Gap rate in data from aggregator	The number of missing measurements or timestamps in the aggregator data per unit time, expressed in gaps/second, gaps/day, or gaps/year. The *gap rate in data from aggregator* can be increased by any communications network and any additional aggregators through which it travels.†*		Y	Y	Y
Attributes of datasets / data completeness / gap size†*					
Mean length of gaps in data from PMU	The mean of the sizes of the known gaps in the PMU data, expressed in units of seconds. The *mean length of gaps in data from PMU* can be nonzero at the PMU, increased at any aggregator, and increased by any communications network through which it travels.†*	An end-user may find the *mean length of gaps in data from PMU* and *mean length of gaps from aggregator* helpful in determining whether the available data will produce suitable results from the desired application. An application may have an **application maximum mean length of gaps**, beyond which the quality of the application's output may degrade. Note that the *application maximum mean length of gaps* will apply to the gap rate of data after it has passed through the entire data process path before the application, not just to the gap rate in data from PMU.	Y	Y	Y
Mean length of gaps from aggregator(s)	The mean of the sizes of the known *gaps* in the aggregator data, expressed in units of seconds. The *mean length of gaps from aggregator* can be increased by any communications network and any additional aggregators through which it travels.†*		Y	Y	Y

Attributes of datasets / data completeness / gap rate *(A gap is defined as the time from when measurements stop until measurements start again. A gap refers to data that is missing completely, not delayed in delivery due to communication drop-outs.)*†*			**Attribute affected by**		
Name of single data attribute	**Metric / specification**	**Related application requirements**	**The PMU?**	**Communications networks?**	**Aggregators?**
Attributes of datasets / data completeness / largest known gap†*					
Largest known gap from PMU	The size of the largest known gap in the PMU data in units of time, expressed in seconds. The *largest known gap from PMU* can be nonzero at the PMU, increased at any aggregator, and increased by any communications network through which it travels.†*	An application may have an **application maximum gap size** in the data which it can tolerate before ceasing to function. For some applications, this may be zero – the application may not tolerate gaps without producing an error. Note that the *application maximum gap size* will apply to the gap rate of data after it has passed through the entire data process path before the application, not just to the gap rate in data from PMU.	Y	Y	Y
Largest known gap from aggregator	The size of the largest known gap in the aggregator data in units of time, expressed of seconds. The *largest known gap from aggregator* can be increased by any communications network and any additional aggregators through which it travels.†*		Y	Y	Y

* Assess regularly – this attribute does not need to accompany every measurement, but should be weighed when determining the suitability of data for an application or study or when troubleshooting problems with application results.

† There are a number of statistics that could be used to characterize a dataset or an incoming PMU data stream; this attribute is provided as one example. Which statistics should be chosen to consider the fitness-for-use of a dataset or an incoming PMU data stream by an application must be determined by the user for the application and specific study to be performed.

Attributes of datasets / characteristics of the data process path *(Sound business practices and explicit quality of service requirements are essential for attaining and maintaining the data process path and should be considered before procuring any element)**			Attribute affected by		
Name of single data attribute	**Metric / specification**	**Related application requirements**	**The PMU?**	**Communications networks?**	**Aggregators?**
Dataset manageability	The ability to create and maintain an effective quality of service that delivers the PMU dataset to the application in a timely manner. For batch processing of a static dataset, if the application is run once per year, a manageable delivery process could involve in-person retrieval of data from the PMU site. If the application is run once per hour, some faster form of data delivery is required for the data process path to be manageable.*		Y	Y	Y
Dataset recoverability	The ability to re-establish delivery of the PMU data to the application after an interruption. For *datasets*, this recoverability can extend beyond just the network communications path being restored. It can also include retrieval from remote data archives manually (e.g., fault data recorders) or assembling from off-site archives of the same data. This approach indicates there are no real-time constraints for the recoverability; those constraints are separated into the incoming PMU data stream attributes of the next section.*		Y	Y	Y
Dataset reliability	The consistency with which data is delivered to the application in a timely manner. For batch processing of a static dataset, there may be multiple options to maintain such timely delivery, such as using backup methods of retrieval from the PMU or aggregator.*		Y	Y	Y
Dataset serviceability	The ability to determine the existence of problems in the PMU data delivery, diagnose their cause(s), and repair those problems.*		Y	Y	Y

* Assess regularly – this attribute does not need to accompany every measurement, but should be weighed when determining the suitability of data for an application or study or when troubleshooting problems with application results.

Attributes of an incoming PMU data stream / characteristics of the data process path (*Sound business practices and explicit quality of service requirements are essential for attaining and maintaining the data process path and should be considered before procuring any element thereof*)*			**Attribute affected by**		
Name of single data attribute	**Metric / specification**	**Related application requirements**	**The PMU?**	**Communications networks?**	**Aggregators?**
Data stream manageability	The ability to create and maintain an effective quality of service that delivers the incoming PMU data stream to the application within the intended *message delivery time,* without exceeding the expected drop-out rate and the expected drop-out size. This includes but is not limited to an effective and sufficient communications network.*	See *message delivery time* below.	Y	Y	Y
Data stream recoverability	The ability to re-establish delivery of the incoming PMU data stream to the application after an interruption, including the ability to restore the stream to delivering data without exceeding the *expected drop-out rate* and the *expected drop-out size.**		Y	Y	Y
Data stream reliability	The ability to deliver the incoming PMU data stream to the application at a specified message rate for a stated period, such as the desired *message continuity period.**		Y	Y	Y
Data stream serviceability	The ability to determine the existence of problems in the incoming PMU data stream delivery, diagnose their cause(s), and repair those problems. An application's requirements for serviceability are largely covered under its requirements for reliability. However, because high-reliability communications networks are not cost-free, an organization may need to define both reliability and serviceability requirements for its PMU data stream that fit the organization's resources.*		Y	Y	Y

* Assess regularly – this attribute does not need to accompany every measurement, but should be weighed when determining the suitability of data for an application or study or when troubleshooting problems with application results.

Attributes of an incoming PMU data stream / data stream availability (*Communication drop-outs are distinct from gaps in the data where the data is missing entirely, not just delayed in communication. Drop-out rate is defined as the total number of messages lost in a time period, whether they are lost consecutively or not.*)*****			**Attribute affected by**		
Name of single data attribute	**Metric / specification**	**Related application requirements**	**The PMU?**	**Communications networks?**	**Aggregators?**
Message rate	The expected number of message reports per second for an incoming PMU data stream, expressed in messages per second. The *message rate* of an incoming PMU data stream is distinct from the *PMU reporting rate;* the *PMU reporting rate* is how often a given PMU is set to take a measurement, while the *message rate* of an incoming PMU data stream has to do with how often the stream is expected to deliver a new message. Realities of measurement and network transport mean that these two rates can and do differ, particularly when there is a problem.*	There may be an **application minimum message rate** defining the lowest message rate at which the application is designed to function. There may be an **application maximum message rate** defining the greatest message rate an application can absorb. The *application maximum message rate* may be defined differently for the same applications running on different platforms.	Y	Y	Y
Message rate type	Whether or not the *message rate* of an incoming PMU data stream is expected to be *fixed* or *variable.**	An application may require that the incoming PMU data stream deliver a fixed *message rate.*	Y	Y	Y
Message arrival order correctness	Y/N: Whether or not the messages of a data stream arrive in the correct order, i.e., the order of increasing timestamps.*	An application may require *message arrival order correctness* for its input stream. Note that this might require a preprocessor sorting the messages into the correct order.	Y	Y	Y
Message delivery time	The time between the timestamp on a measurement(s) in a message and the arrival of the message at the application, expressed in seconds to a specified number of decimal places. Network performance, particularly latency, may be the most important factor affecting *message delivery time.**	An application may have an **application maximum message delivery time,** which is the *measurement delivery time* at which point the application can no longer deliver results within an operationally useful timeframe.	Y	Y	Y

* Assess regularly – this attribute does not need to accompany every measurement, but should be weighed when determining the suitability of data for an application or study or when troubleshooting problems with application results.

Attributes of an incoming PMU data stream / data stream availability (*Communication drop-outs are distinct from gaps in the data where the data is missing entirely, not just delayed in communication. Drop-out rate is defined as the total number of messages lost in a time period, whether they are lost consecutively or not.*)*****			**Attribute affected by**		
Name of single data attribute	**Metric / specification**	**Related application requirements**	**The PMU?**	**Communications networks?**	**Aggregators?**
Expected drop-out rate	The rate at which messages may be lost from the *incoming PMU data stream*, while the *PMU data stream* is still considered live and operational. Expressed in messages per second, messages per minute, or messages per day. Network performance, particularly latency and jitter, may be the most important factor affecting *drop-out rate.*†*	An application may have an **application maximum drop-out rate,** which is the *drop-out rate* at which point the application can no longer deliver operationally useful results, where *drop-out rate* is defined as the total number of messages lost in a time period, whether they are lost consecutively or not.	Y	Y	Y
Expected drop-out size	The number of messages in a row that may be missing while the *PMU data stream* is still considered live and operational. Network performance, particularly latency, may be the most important factor affecting *expected drop-out size*.†*	An application may have an **application maximum drop-out size,** which is the *drop-out size* at which point the application can no longer deliver operationally useful results.	Y	Y	Y

* Assess regularly – this attribute does not need to accompany every measurement, but should be weighed when determining the suitability of data for an application or study or when troubleshooting problems with application results.

† There are a number of statistics that could be used to characterize a dataset or an incoming PMU data stream; this attribute is provided as one example. Which statistics should be chosen to consider the fitness-for-use of a dataset or an incoming PMU data stream by an application must be determined by the user for the application and specific study to be performed.

Attributes of an incoming PMU data stream / data stream availability (*Communication drop-outs are distinct from gaps in the data where the data is missing entirely, not just delayed in communication. Drop-out rate is defined as the total number of messages lost in a time period, whether they are lost consecutively or not.*)*****			**Attribute affected by**		
Name of single data attribute	**Metric / specification**	**Related application requirements**	**The PMU?**	**Communications networks?**	**Aggregators?**
Message continuity period	A period of time for which the *message rate* remains as expected, with no more *drop-outs* than the *expected drop-out rate* and with no *drop-out* exceeding the *expected drop-out size,* expressed in days:hours:minutes:seconds to a specified number of decimal places. Synchrophasor network performance, particularly latency, may be the most important factor affecting the *message continuity period*.†*	An application may have a specified **application time window** defining the minimum period of data retention for the application to process an incoming data stream. The *application time window* is defined as a period preceding the current time during which the *application maximum drop-out size* and *application maximum drop-out rate* must not be exceeded. This is distinct from the question of *application training period*; an application might remain adequately "trained" to resume analysis without loss of methodological soundness despite an hour's interruption, but its processing of an incoming stream might require five minutes of message stream preceding the current time interval to start producing a new stream of results.	Y	Y	Y

* Assess regularly – this attribute does not need to accompany every measurement, but should be weighed when determining the suitability of data for an application or study or when troubleshooting problems with application results.

† There are a number of statistics that could be used to characterize a dataset or an incoming PMU data stream; this attribute is provided as one example. Which statistics should be chosen to consider the fitness-for-use of a dataset or an incoming PMU data stream by an application must be determined by the user for the application and specific study to be performed.

Appendix B – Brief Review of Data Quality and Accuracy History

At the start of this effort, the PARTF whitepaper team conducted extensive research into the definition and categorization of data quality and availability within the synchrophasor field and in other fields that 1) have substantial areas of technical overlap with the field of synchrophasor data generation, transport, storage, and its many uses; 2) have a strong consensus in the community that developed those definitions on what they should mean; and 3) had a history of spirited debate before arriving at that consensus. Sets of terms and definitions were desired from communities of experts that had put in some of the work for us wrangling out the details of what needed to be defined. Technical areas that had definition sets that matched these criteria included information technology, geographic information systems (GIS), and telecommunications.

A key lesson, learned from examining existing definition sets and through spirited debates that eventually produced a consensus, was that when experts on a topic find it difficult to agree on the meaning of any one term, very often the underlying problem is that the term in question is actually an umbrella term for multiple attributes that each need their own name and definition.

The GIS field has a long history of discussion regarding meanings of accuracy and data quality with respect to sets of data that involve both *measurement accuracy* and attached *attribute accuracy* due to the crucial nature of attached Global Positioning System (GPS) stamps.

Data quality is an umbrella term for the many attributes of a gathered dataset that make it fit or unfit to be used for certain tasks and define its overall integrity or lack thereof. What those attributes are and how they are defined are complicated questions with answers that vary according to the particular problem and situation. One possible approach to beginning to define *data quality* is given by the Spatial Data Transfer Standard (SDTS) (USGS 1998), a data transfer standard designed to facilitate dissemination and sharing of data. SDTS provides standard definitions of data elements, a standardized format for data transfer, and descriptive metadata for database content. SDTS was adopted by the National Institute of Standards and Technology as a Federal Information Processing Standard (FIPS-173) in 1992 and withdrawn in 2005. While no longer active, its original components of data quality still serve as a starting point for minimum attributes to specify data quality in a given field; however, these components are only a starting point and additional attributes must also be defined. The five components of *data quality* identified in SDTS include the following:

- data lineage
- positional accuracy
- attribute accuracy
- logical consistency
- data completeness.

These five primary components, and related source materials of the attributes that make up these components, were used to inform the discussion presented in this report. The top-level categorizations and the usage of terms in this report differ substantially and many additional attributes have been added to tailor the definition set in this report to the unique needs of the synchrophasor data application field.

The body of literature associated with GIS data quality and SDTS may prove useful for further reading to inform application requirements for a particular study or other topics related to data quality and accuracy. For example, while the scope of this work includes describing data as it is, there is substantial discussion in the GIS literature on what is necessary to produce true data accuracy, such as the discussions on *model* *completeness*. Recommended related reading includes USGS (1998), Guptill (1995), Buckley (1997), NASA (1994), and Modell (1994).

Sources and recommended reading from the synchrophasor field include: Liu (2015), Meliopoulo (2007), Huang (2008), Kirkham (2015), Hu (2010), Hasan (2009), and Goldstein (2015).

Recommended further reading on information technology and telecommunications: Kihl (2011) and Cavin (2011).

"Actual" value, _real value_, or other representations of "truth"

This document uses "the signal being measured" to indicate what is intended to be measured, contrasting with the "measured result" which will usually have some level of error. Historical metrology sources often use phrases such as "true value" without ever defining what that means, trusting the reader to get the idea. Over time using any form of "true" has become heavily disfavored in metrology circles because defining what truth is becomes problematic. "Real value" has been suggested, but in a discussion that includes complex numbers, confusing the meaning of "real" is not helpful.

Several terms to indicate the grid condition that is intended to be measured were tried out in early drafts of this document, including:

- *Realized quantity*: This was not sufficiently intuitive; readers remarked that they had to remind themselves what it meant.

- "*Actual" value*: This was found to be intuitive, but awkward and still likely to generate arguments as to what it means.

Exploring deeper meanings of truth in metrology is outside the scope of this document. That being the case, this document uses "the signal being measured" to indicate the grid condition that is intended to be measured in order to avoid such arguments.

Other work is currently going forward using "ideal measured value" to indicate the measurement that is ideally wanted.