• What is the context/background?

Traditionally wind sites have collected real-time data on locally hosted databases. The historical data would be used to trend site performance, build business models, perform audits and more. The data would be stored in aggregate form to reduce the amount of data collected, typically because of the constraints of the computer hosting the database. Often 10-minute records would suffice for any historical data. Today, more data is requested from these systems to support or implement technologies such as Artificial Intelligence (AI) [1]. The growth of cloud storage, services and IoT technologies has decreased the dependency on hosted databases and older protocols, allowing for more granular data to be captured and processed at higher rates. The amounts of varying and variable sensor data collected can be categorized as “Big Data”. The data collected is analyzed to reveal the knowledge of unseen patterns as described by Shyam. R [2].

In order for analysts and decision makers to produce accurate analysis, make effective decisions and take actions, data must be trustworthy. It is thus important to provide a comprehensive solution for assessing and assuring the trustworthiness of the information collected. In general, the problem of providing “good” data to users and applications is an inherently difficult problem which often depends on the application and data semantics as well as on the current context and situation. In many cases, it is crucial to provide users and applications not only with the needed data, but with also an evaluation indicating how much the data can be trusted. Bertino and Lim [3] proposed computational methodologies to generate trust scores of every element of data in the context of distributed systems.

It is thus important to provide a comprehensive solution for assessing and assuring the trustworthiness of the information collected in such data sharing systems since decisions and analyses are largely affected by this information. Attacks or unexpected accidents may result in bad data being provided to critical components of the system. These components may in turn take wrong decisions or generate inaccurate analyses that can result in damages to real-world objects such as manufacturing facilities or power plants. For example, in an electric power grid which consists of 270 utilities using a SCADA (Supervisory Control and Data Acquisition) system that can contain up to 50,000 data collection points and over 3,000 public/private electric utilities, any single point of failure can disrupt the entire process flow and can potentially cause a domino effect that shuts down the entire systems [11]

Machine learning from large data sets can provide new insights to the operations of wind sites. Wind sites need to collect more data than ever. In order to grade the quality of the data collected, domain specific rules can be designed to measure different dimensions that can be used by data scientists and machine learning projects.

• What is the problem?

When collecting large datasets from wind sites, rules need to be defined in order to describe the quality of the data before its used in modeling and real time operations. Defining these rules can be arbitrary or specific to the end user. A set of general methods can be proposed if the domain and end-user is known. Given a typical wind site architecture and assuming that the end user of the data will be data scientists and machine learning algorithms , the authors attempt to present methods for data quality metrics.

• What is the state of the art?

The existing literature on defining data quality rules provided insights on what are useful dimensions to measure. Reviewing the metrics most widely used in wind site anomaly-based detection literature, data quality dimensions where selected that would most benefit wind site data. Connectivity, Completeness and Consistency are metrics that can be defined directly using industry standards or arbitrary business rules. The dimensions of Accuracy and Timeliness are analyzed deeper and wind-site domain specific methods are proposed to provide these metrics.



• What is the proposed solution?

Research effective ways to implement data quality rules for each dimension and test the rules on a real wind farm.

Wind site that the rules will be tested on :

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| # WTGs | Site MW | Base Rating | Rotor | Blade Types | Pitch Type | Tower Height | Controller | Converter Type | Software Version | SCADA Version |
| 37 | 59.9 | 1.62 | 100 | Glass | Salem | 80 | MarkVI | GE ESS | 44.76.00C | WindSCADA 10.0 SP2 |

Connectivity – OPC error codes , industry protocol standard…

Consistency – Using server uptime metrics on connectivity to define …

Completeness – using IEC standard heierachys for wind turbines critical tag sets can be created and meas

• What is the performance?

Connectivty –

* no performance metric? OPC error codes measured,
* rarely anything other than “bad”
* Issues injected , to verify the conectivty works.
* A strong indicator of connectivity

Consisitency -

* Server uptime , measured using a heart beat metric
* Inversely measuring the value from the collection system allowed the measurement of the system when it was unavailable.
* Strong metric for measuring

Completeness

* IEC Standard was used to create a set for operations
* This can be modified for other business customers
* A list of critical tags was used to validate,
* Reports showe that tags from similar turbines would be missing
* This resulted in a secondary report that showed that 5-10% of turbines at a site would have missing tags.
* Missing tags where found to be related to inconsistent programming of turbine controllers based on software revisions or major changes
* The metric is too specific and a more general method may be identifying turbines that are not the same as majority. Clustering tag anmes could do this

Timeliness

* Using the proposed method, drifts in the data could be detected using secondary and tertiary channels at the windsite to verify power. The drifts where measured over the course of X time

Accuracy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tag** | **Average** | **Max** | **Min** | **Standard Deviation** |
| In\_WindDirection | 45.2 | 48 | 33 | 2.4 |
| AI\_CuReacPowerAct | 46.7 | 48 | 32 | 2.4 |
| AI\_In\_GridMonRealPowerAct | 46.7 | 48 | 33 | 2.4 |
| In\_WindSpd | 45.2 | 48 | 32 | 2.6 |
| AI\_In\_GridMonAppPowerAct | 46.7 | 48 | 33 | 2.3 |
| In\_RotorSpd | 44.65 | 48 | 4 | 8.2 |
| AI\_In\_GridMonReacPowerAct | 46.7 | 48 | 33 | 2.3 |
| AI\_In\_GridMonCosPhiAct | 46.6 | 48 | 32 | 2.4 |
| OpCtl\_PowerSetpt | 41.9 | 48 | 1 | 13.1 |

“The authors calculated the detection accuracy as the metrics of performance… Compared with methods A & B,

the proposed method achieved…”

• What is the conclusion?

“The authors proposed an anomaly-based detector using … The method achieved an accuracy of … Possible

future improvements include …”

Problem:

When multiple wind sites are managed the local historical databases can be released by historian solutions such as OSI PI. These historian solutions provide real time data collection into compressed databases that can be hosted remotely in a central location. The historian database solution has been widely implemented. It allows for interpolation of the raw data and provides many configuration capabilities. It also allows for historical data browsing through many business tools such as excel plugins, and (software development kits) SDK’s. The cost of these systems are dependent on licensing fees and the infrastructure of the database servers that are hosted on-premise.

Two critical technologies have emerged in recent years that is challenging the current paradigm of data collection from wind sites. First, cloud technologies reduce the overhead of hosting data centers and opens the ability for users to leverage elastic computing on the data hosted on the cloud. Traditionally business users needed to offload data from a database to generate insights but now much more powerful tools are located alongside the data and only require the user to interact with it. Secondly, machine learning has risen as a tool that can generate deeper insights on large datasets reducing the need of traditional business analyst roles. It can also provide insights and solutions for controls, performance and automation in conjunction with edge computing.

Many cloud-based historian and SCADA solutions have been emerging to help the transition into this new paradigm. The main focus of these new collection strategies is to increase the data collected locally using new configurations of the real-time data. This is possible because all of the data collected is now hosted on affordable and scalable storage that has been tuned for time-series data streams. The querying of the cloud storage is done by secondary systems allowing for further improvement of the data ingestion. With greater capacity for data ingestion more sensors, auxiliary channels, and secondary systems can now be added to the dataset of a wind site. With more adhoc data available the traditional documentation of the data may not be accessible or simply unavailable. This paper leaves the details of these solutions for the review of the reader.

The method used to add additional data to a wind site was to systematically document the source or purpose of the measurement and integrate it on the collection client at the wind site. It allows for simple analysis by the business users. This method may no longer be possible with the amount of adhoc data collected. Explicit descriptions may not be needed when the data is analyzed by machine learning algorithms that can cluster and identify characteristics. Instead of descriptions what is needed are well defined quality rules to assist in data cleansing.

The paper explores how to define five arbitrary data quality dimensions such as Connectivity, Completeness, Timeliness, Accuracy, and Consistency. There are many more of these dimensions defined in other domains, but for the purpose of this paper these concepts will be reviewed as they apply to a wind site and its associated wind turbines.

By analyzing the several wind sites that use GE technologies the paper explores how to define

Hypotheses tested :

1. Chosen definitions are adequate for data quality rules.
2. Assumptions are set for the typical wind site architecture
   1. Network topology
   2. GE technology
   3. Windsite with multiple wind turbines
   4. Use of OPC standard for communications
   5. Use of cloud computing and adhoc data channels
   6. Some knowledge of wind site data not complete
   7. Power and wind speed data channels readily available
3. Algorithms used are statistical steps to measure resolution and benchmark the thresholds needed at each wind site.
4. Analysis of wind turbine channels of data can be analyzed using python scripts
   1. Data completeness was measure by running ana analysis on several wind sites for critical tags that hav been identified for multiple users. IEC standard names can be used to heirachally identify the critical tags that are required for the completeness measurement.

Results:

1. Accuracy dimension was identified
   1. Resolution of same data chanels seems to vary between sites
   2. Even through collection systems can be configured to capture as fast as possible (1sec) , due to network congestion, protocol limitations, and possible software latencies the resolution of data channels can fluctuate over time.
   3. Packet transmissions where found to be 1280 ms through wireshark captures at some sites while other sites the communications where found to be closer to 1000 ms.
   4. Points per minute of a data channel can become a metric that can be used for SCADA anaomoly detection
   5. Capturing the points per minute in ageneric method was proposed
      1. Initially sort the chennels by frequency and create a baseline of ppm for the highest frequency channels.
      2. Capture the measurements in a statistically significant method.
      3. A 24 hour sample size with 100 samples of 1 mniute measurements allowed for less than 10% margin of error.

Findings:

1. Accuracy DQ can be used for anamoly detection and forestablishing a quality metric
   1. Can be used to filter out data that did not arrive with high resolution to build better models off high res data
   2. A formula that has can leerage a running metric could be a better approach
      1. With a running metric any deviation from the previous days ppm can be used for assigning a resolution quality metric

what data quality rules can be applied to the collection system of wind site data set for the eventual consumption of users interested in data science from the data. Requirements of what the end user needs can vary so the focus shall be on defining generic methods for capturing metrics that can be ued by end users as a foundation for assessing the health of the data for specific use cases and analysis.

There are unique conditions at a wind site compared to a solar, nuclear or fossil site that can be used to define data quality rules. The paper uses a typical architecture of a wind site

# Defining Data Quality Dimensions

Intrinsic DQ dimensions can be directly measured from data collection system

Contextual DQ Dimensions require insights, benchmarks, analysis and historical data to measure

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data Quality Dimension** | **Class** | **Properties** | **Description** | **Wind Site Data Definition** |
| Connectivity | Intrinsic | Infrastructure, Application, Site | Measure of source availability while collecting data | Defined as ‘Good’ or ‘Bad’ at different levels of operations  A channel of data is not connected when the channel error code is defined as ‘Bad’.  A wind turbine is not connected when any of the ‘critical’ channels are defined as ‘Bad’. |
| Completeness | Intrinsic, Contextual | Critical Data, Dark Data, Data Leakage | Measure of data collected from a specified data set | The percentage of channels that exist at a wind turbine that are required by a customer |
| Timeliness | Intrinsic | Processing Latency, Synchronization | Data synchronized across all sources in spite of time stamp capture or system error | Latency of real power, measured in milli-seconds between SCADA channels and site power meters  Latency of average wind turbine windspeed measurement and site met towers |
| Accuracy | Intrinsic, Contextual | Sampling Resolution, Precision, Filtering | Measure of how accurate the data captured represents the generated data (in time and value) | Given the top 10% high frequency channels at a wind turbine measure the percent deviation between the historical resolution and the sampled average resolution.  Given the top 10% high precision channels at a wind turbine measure the percent deviation between the historical precision and the sampled average precision.  Take max percentage deviation from either metrics mentioned above to create an Accuracy metric as a percentage. Where 100% means no significant change between historical resolution or precision and sampled resolution or precision |
| Consistency | Contextual | Batching, Smooth Variance, Presentation | Measure of data collected from a specified data set over time. | Uptime percentage measured in number of minutes the system is operational over the course of a year. |

# DQ Connectivity

Connectivity of a point of data would always be a Boolean of “connected” or “not connected”, but in order to attribute a quality metric to the “not-connected” state a deeper look into the connection failure modes is required.

The data is collected from the wind sites Supervisory Control and Data Acquisition (SCADA) system. The SCADA system is typically a physical or virtual server that hosts the manufacturers proprietary SCADA software that amongst other things, manages the collection of data from each wind turbine.



## Classification

Consider the proposed arbitrary classification for an item that is not connected, “Run-Time” and “Configuration”. Where “Run-Time” errors are errors that occur while a connection is established, and where “Configuration” errors are errors that occur when a connection is attempting to be established. These classes can apply to many SCADA system regardless of the technology where the specific error codes could be different.

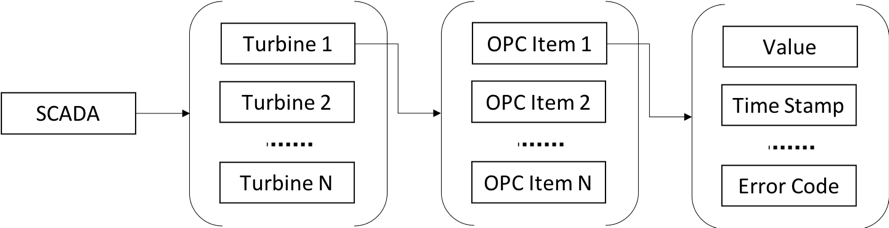
## Severity

The severity of the error can range from a recommendation to a fatal error preventing all data from working. A proposed method is to further extract severity level for each error code this method is generic to all SCADA systems and can be used to generalize error codes to enhance the Connectivity metric for each wind turbine. A common error logging standard was chosen to indicate the severity of each error code. The severity levels are defined as follows:

* Warning : Partial data loss to a single item, or indicator of sub-optimal performance
* Error : Loss of all data from a single item
* Fatal : Loss of data from many or all items

## Implementation

A Wind farm comprised of General Electric Wind Turbine technology will typically provide its GE WindSCADA software that publishes the wind turbine data over an OPC Server. The OPC Server provides channels of data called “Items”. These OPC Items contain an error-code property that has a description of the specific error. The OPC Error Codes are published in the GE WindSCADA Manual (GE).

****

| GE WindSCADA OPC *Error Codes* | | DQ Error Code Classification | |
| --- | --- | --- | --- |
| ***OPC Error Code*** | ***Description*** | ***Error Code Class*** | ***Error Class Level*** |
| OPC\_E\_BADTYPE | The passed data type can not be accepted for this item from server | Configuration | Error |
| OPC\_E\_BADRIGHTS | The Item is not having either Readable or writable access rights | Configuration | Error |
| OPC\_E\_RANGE | The value was out of Range | Run-Time | Warning |
| OPC\_E\_INVALIDHANDLE | Clients Item handle is invalid when requested to server | Configuration | Fatal |
| E\_NOINTERFACE | The possible version conflict between the OPC DA server version and OPC Client version while communicating | Configuration | Fatal |
| OPC\_E\_UNKNOWNITEMID | The Item ID is not part of OPC Namespace in the OPCDA server | Configuration | Error |
| OPC\_E\_INVALIDITEMID | The client requested item Name has invalid convention (for ex some invalid characters) | Configuration | Warning |
| OPC\_E\_DUPLICATENAME | Trying to add a group which is already present in server | Configuration | Warning |
| OPC\_E\_NOTSUPPORTED | If a Client attempts to write any value, quality,timestamp combination and the server does not support the requested combination(which could be a single quantity such as just timestamp), then the server will not perform any write and will return this error code | Run-Time | Warning |
| E\_OUTOFMEMORY | Not Enough memory to complete the requested operation. This can happen any time the server needs to allocate memory to complete the requested operation | Run-Time | Fatal |
| E\_FAIL |  | Configuration/Run-Time | Fatal |
| OPC\_S\_CLAMP | The Value was accepted but was clamped | Run-Time | Warning |
| E\_INVALIDARG | An invalid argument was passed(like when client requests data to server the argument of dwcount should be >0 but if dwcount=0 then this error code will be returned | Run-Time | Warning |
| CONNECTION\_E\_CONNECTIONT | The client has not registered it communication channel with server for the data updation | Configuration | Error |
| OPC\_E\_DEADBANDNOTSUPPORTED | The dead band is not supported by the server | Configuration | Warning |
| OPC\_S\_UNSUPPORTABLERATE | Server does not support requested rate,server returns the rate that it can support in the revised sampling rate | Run-Time | Warning |
| OPC\_E\_NOBUFFERING | The server does not support buffering of data items that are collected at a faster rate than a group update rate | Configuration | Warning |
| OPC\_E\_UNKNOWNPATH | The Item’s access path is not known to the server | Configuration | Error |
| OPC\_S\_INUSE | The operation cannot be performed because the object is being referenced | Run-Time | Error |

* GE OPC Error Codes provide detailed connection error descriptions
* Error Code Classifications can be defined as “Configuration”, and “Runtime”

Error Codes are captured using special property ID’s defined in GE Manual

|  |  |
| --- | --- |
| **Data Quality Dimension** | **Identification Method** |
| Connectivity | * GE OPC Error Codes provide detailed connection error descriptions * Error Code Classifications can be defined as “Configuration”, and “Runtime” * Error Codes are captured using special property ID’s defined in GE Manual |
| Completeness | * Assumption for Data Completeness is “Wind Farm Auto-Correlation”, All turbines should have the same amount of tags at a site. * Use of IEC 64100-25 Standard Logical Node Hierarchy identifies min. viable tags per wind turbine ‘LN’ grouping. * Crititcal tag list by customer |
| Timeliness | **Time Stamp Synchronization / Drifts:**   * NTP server synchronization, Conversion to UTC from local time stamps, provide data to user as local time from a UTC timestamp * The timestamp of the data when it was captured * The timestamp of the data when it is recorded * The timestamp of the data when it is read by the user * If the system has ‘n’ number of timestamps from measurement to final reading then all-time stamps should remain in chronological order of capture. * Include any known process delay time to account for drifts in server synchronizations.   **Measuring Data Latency**   * **Capture real time power value from a substation source and comparing to the power of the scada to** * **4 seconds of latency** * **Comparing latency of measurements with other measurements sources.** * **Identified SCADA power measurments is a separate source than the meter measurements** * **Using the two measurements any latencies can be detected between sources.** * **Multiple meters can be used to measure the output of the scada system .** * **Multiple metereological measurements from the wind turbines can be captured and measured against the site met towers to confirm latency anamolies.** |
| Accuracy | **Resolution:**   * Resolution metric must be identified statistically across many sites by benchmarking against high frequency channels/tags of a high-res site. * Take 100 discrete uniformly distributed values of the resolution metric (points/min) of a high frequency channel over the course of 24 hours. * After characterizing a channel compare it to other turbines at the site to get a site level frequency range average with standard deviation. * Set a tolerance for the channel frequency benchmark and compare deviation from norm when grading similar channel/tag at other sites. * Choose top 10 channels with highest frequency for benchmarking.   **Precision:**   * Quantization errors ,Truncation , Signed/Unsigned integers * Scaled floats into ints * Engineering Unit Max and Min limits exceeded (Provided by GE OPC Item property) * 64-bit data type on driver but reading 32 bit would cause precision losses * Reading 64-bit datatype given 32-bit value would cause random values at high significant digits |
| Consistency | **Historical Analysis of Other DQ dimensions**   * Change in configuration settings * Change in Accuracy over time : Resolution/Precision * Change in Connectivity over time * Change in Completeness over time * Change in Timeliness over time |

# References

Effects of the pre-processing algorithms in fault diagnosis of wind turbines

Pere Marti-Puiga,∗,1, Alejandro Blanco-Ma,b,1, Juan José Cárdenasb, Jordi Cusidób, Jordi Solé-Casalsa

Comparison of API & IEC Standards for Offshore Wind Turbine Applications in the U.S. Atlantic Ocean: Phase II March 9, 2009 – September 9, 2009 A. Jha, D. Dolan, T. Gur, S. Soyoz, and C. Alpdogan MMI Engineering Houston, Texas NREL Technical Monitor: Walt Musial

1. Aimad Karkouch “Data Quality in Internet of Things: A state of the art servey” 2016

2. [Sabrina Sicaria](https://www.sciencedirect.com/science/article/pii/S0306437916300072" \l "!), [Alessandra R](https://www.sciencedirect.com/science/article/pii/S0306437916300072#!) “A secure and quality-aware prototypical architecture for the Internet of Things” 2016

2. Zita Vale,Hugo Morais,Pedro Faria,Carlos Ramos Distribution system operation supported by contextual energy resource management

based on intelligent SCADA , GECAD – Knowledge Engineering and Decision Support

Research Center, Polytechnic of Porto (IPP), R. Dr. António Bernardino de Almeida, 431, 4200-072 Porto, Portugal

3. K.Metaxiotis A.Kagiannas D.Askounis J.Psarras , Artificial intelligence in short term

electric load forecasting: a state-of-the-art survey for the researcher

Department of Electrical and Computer Engineering, National Technical University

Athens, 9 Iroon Polytechniou str., Zografou 15773, Athens, Greece

Received 12 February 2002, Accepted 10 June 2002, Available online 15 October 2002.