Interdisciplinary Journal of Information, Knowledge, and Management

An Official Publication of the Informing Science Institute
InformingScience.org

IJIKM.org

Volume 14, 2019

A NEW TYPOLOGY DESIGN OF PERFORMANCE METRICS TO MEASURE ERRORS IN MACHINE LEARNING REGRESSION ALGORITHMS

Alexei Botchkarev — GS Research & Consulting and Ryerson University.Toronto, Canada — alex.bot@gsrc.ca

ABSTRACT

Aim/Purpose	The aim of this study was to analyze various performance metrics and approaches to their classification. The main goal of the study was to develop a new typology that will help to advance knowledge of metrics and facilitate their use in machine learning regression algorithms
Background	Performance metrics (error measures) are vital components of the evaluation frameworks in various fields. A performance metric can be defined as a logical and mathematical construct designed to measure how close are the actual results from what has been expected or predicted. A vast variety of performance metrics have been described in academic literature. The most commonly mentioned metrics in research studies are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), etc. Knowledge about metrics properties needs to be systematized to simplify the design and use of the metrics.
Methodology	A qualitative study was conducted to achieve the objectives of identifying related peer-reviewed research studies, literature reviews, critical thinking and inductive reasoning.
Contribution	The main contribution of this paper is in ordering knowledge of performance metrics and enhancing understanding of their structure and properties by proposing a new typology, generic primary metrics mathematical formula and a visualization chart

Accepting Editor Felix O Quayson | Received: September 11, 2018 | Revised: November 15, December 20, 2018; January 13, 2019 | Accepted: January 17, 2019.
Cite as: Botchkarev, A. (2019). A new typology design of performance metrics to measure errors in machine learning regression algorithms. *Interdisciplinary Journal of Information, Knowledge, and Management*, *14,* 45-79. https://doi.org/10.28945/4184

Findings	Based on the analysis of the structure of numerous performance metrics, we proposed a framework of metrics which includes four (4) categories: primary metrics, extended metrics, composite metrics, and hybrid sets of metrics. The paper identified three (3) key components (dimensions) that determine the structure and properties of primary metrics: method of determining point distance, method of normalization, method of aggregation of point distances over a data set. For each component, implementation options have been identified. The suggested new typology has been shown to cover a total of over 40 commonly used primary metrics
Recommendations for Practitioners	Presented findings can be used to facilitate teaching performance metrics to university students and expedite metrics selection and implementation processes for practitioners
Recommendations for Researchers	By using the proposed typology, researchers can streamline development of new metrics with predetermined properties
Impact on Society	The outcomes of this study could be used for improving evaluation results in machine learning regression, forecasting and prognostics with direct or indirect positive impacts on innovation and productivity in a societal sense
Future Research	Future research is needed to examine the properties of the extended metrics, composite metrics, and hybrid sets of metrics. Empirical study of the metrics is needed using R Studio or Azure Machine Learning Studio, to find associations between the properties of primary metrics and their "numerical" behavior in a wide spectrum of data characteristics and business or research requirements
Keywords	performance metrics, error measures, accuracy measures, distance, similarity, dissimilarity, properties, typology, classification, machine learning, regression, forecasting, prognostics, prediction, evaluation, estimation, modeling

INTRODUCTION

Performance evaluation is an interdisciplinary research problem. Performance metrics (error measures) are vital components of the evaluation frameworks in various fields. A performance metric can be defined as a logical and mathematical construct designed to measure how close are the actual results from what has been expected or predicted. Among most commonly used Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) can be mentioned. In a generic sense, performance metrics are linked to the scientific concepts of distance and similarity. In machine learning regression experiments, performance metrics are used to compare the trained model predictions with the actual (observed) data from the testing data set (e.g., Botchkarev, 2018a; Makridakis, Spiliotis and Assimakopoulos, 2018). The results of these comparisons can directly influence the decision-making process of selecting the types of machine learning algorithms for implementation.

Deza and Deza (2016) indicate that similarity measures are needed in almost all knowledge disciplines. A long-standing interest in performance metrics can be found in forecasting and prognostics. Forecasting has a long history of employing performance metrics to measure how much forecasts deviate from observations in order to assess quality and choose forecasting methods, especially in support of supply chain or predicting workload for software development (e.g., Carbone and Armstrong, 1982; De Gooijer and Hyndman, 2006). Prognostics - an emerging concept in condition-based maintenance (CBM) of critical systems in aerospace, nuclear, medicine, etc. – heavily relies on performance metrics (e.g., Saxena et al, 2008).

Classification is one of the main topics of scientific research (Parrochia, n.d.). Each knowledge domain, as a subject of scientific research, requires classification systems (typology) to structure the

contents in a systematic manner. Categories of the typology are defined based on resemblances (or differences) of items/objects in a specific context. Typologies are helpful in ordering and organizing knowledge, defining the scope and simplifying studies, facilitating information retrieval and detecting duplicative objects (e.g., Gerber, Baskerville & Van der Merwe, 2017). Multiple performance metrics have been designed and described in academic journals (see References). Knowledge about their properties needs to be systematized in a clear way to simplify the design and use of the metrics. Available classifications have certain drawbacks (e.g., Cha, 2007; Makridakis & Hibon, 1995), which are described in the paper.

The intention of this paper was to review existing performance metrics classifications and develop a typology that will help to improve our knowledge and understanding of a variety of metrics and facilitate their use in machine learning regression, forecasting and prognostics.

The rest of the paper is structured as follows. First, we provide a literature review. Thenwe describe methodology of the study. In the next section, we describe a proposed metrics framework, which includes the following categories: primary metrics, extended metrics, composite metrics and hybrid sets of metrics. The main attention and space in this paper is focused on the properties and typology of the primary metrics. The final sections present discussion and conclusions.

METHODOLOGY

Objectives. The first objective of this study was to provide an overview of a variety of the performance metrics and approaches to their classification (grouping/systematization). The main goal of the study was to develop a new typology that will help to advance knowledge of metrics, enhance understanding of their structure and properties, and facilitate their use in machine learning regression, forecasting and prognostics.

Method. Several research methodologies were used to achieve the objectives: identification of related peer-reviewed papers, critical literature review, critical thinking and inductive reasoning. The study has a qualitative nature. The search was conducted in Google Scholar and several databases through the EBSCO integrated search including Health Business Elite, Health Policy Reference Center, BioMed Central, Business Source Complete, MEDLINE Complete, CINAHL Complete, PubMed, The Cochrane Library, etc. Around 500 papers were retrieved and previewed. Over 80 papers (which used qualitative, quantitative and mixed methods of studies) where selected, reviewed in more detail and cited in the paper.

For better understanding by the readers, the proposed typology has been formulated and presented in several ways: mathematical, verbal, visual. First, mathematical expressions are provided for a generalized metrics construct and metrics components. Complexity of this mathematics does not go beyond the second-year computer science program requirements. Second, all mathematics are accompanied by verbal descriptions of their meaning and practical implications in certain scenarios. Finally, the table-format chart presents a typology in a one-page simple and clear way and ties together all metrics components for easier visual comprehension.

Terminology and abbreviations. As this paper covers research in an interdisciplinary area, which is related to machine learning, prognositcs, forecasting, terminology may vary from field to field.

Our main focus is on performance metrics. In literature, many terms are used with close meaning, e.g., measure, distance, similarity, dissimilarity, index, etc.

Different terms are used in literature regarding grouping performance metrics, e.g., classification, taxonomy, etc. In the literature review, we use the terms used by the authors of the papers under consideration. Later in the paper, we refer to our construct as typology.

Multiple performance metrics are considered in the paper. Commonly, we refer to them using abbreviations. A list of all metrics abbreviations mentioned in the paper is provided in Appendix A. Usual-

ly, the first letters in abbreviations use: M for mean (arithmetic), Md for median, GM for geometric mean.

Mathematical definitions of performance metrics are shown in Appendix B. These metrics are implemented in R Studio (e.g., packages MLmetrics, forecast) and in Azure Machine Learning Studio (e.g., Botchkarev, 2018b). Some metrics have alternative definitions. They are listed in Appendix C.

Performance metrics are designed to compare two data sets. We refer to them as actual, $A = (A_1, A_2, \dots, A_j)$, i.e., a data set containing actual values, and predicted, $P = (P_1, P_2, \dots, P_j)$, i.e., a data set containing predicted values. In literature, depending on the research field, actual may be referred to as observed or measured, and predicted may be called forecasted, modeled, simulated, estimated.

LITERATURE REVIEW

The Literature Review section is structured into two parts aligned with the two main topics of the paper. The first part overviews the most common metrics. The second part provides descriptions of known metrics' classifications, their benefits and drawbacks.

PART I: THE MOST COMMON METRICS

A large variety of metrics has been suggested and used in many knowledge areas. Makridakis and Hibon (1995, p. 3) stated that "there are fourteen accuracy measures which can be identified in the forecasting literature". It seems that no other author risked offering an exhaustive list of metrics. Usually, a list of metrics is accompanied with qualifiers: most popular, commonly, widely or frequently used, etc. There are many analytic reviews covering dozens of metrics. Kyriakidis, Kukkonen, Karatzas, Papadourakis and Ware, (2015) studied 24 metrics used in air quality forecasting. De Gooijer and Hyndman (2006), in a review covering 25 years of time series forecasting, list 17 commonly used accuracy measures. Shcherbakov et al (2013) presented a survey of more than twenty forecast error measures. Prasath, Alfeilat, Lasassmeh and Hassanat (2017) studied 54 (fifty-four) measures and their effect on machine learning of K-Nearest Neighbor Classifier (KNN). Numerous distance metrics from diverse knowledge domains are compiled and briefly described in the Encyclopedia of distances (Deza & Deza, 2016).

Some metrics are more popular than the others. Several researchers conducted surveys of organizations and practitioners to understand the frequency of use or importance of different metrics. A variety of metrics were identified in these surveys. However, top most common metrics came up in many studies. Table 1 shows three metrics found most popular in the independent surveys that were performed over a timeline of 25 years: mean square error (MSE) (or root MSE (RMSE)), mean absolute error (MAE) and mean absolute percentage error (MAPE).

Table 1. Top three metrics identified in the surveys, percentage

Metrics	C&A, 1982	M&K, 1995	M et al, 2006	F&G, 2007
Mean square error (MSE) or Root MSE (RMSE)	34	10	6	9
Mean absolute error (MAE)	18	25	20	36
Mean absolute percentage error (MAPE)	15	52	45	44

Note: C&A, 1982 – study by Carbone and Armstrong (1982); M&K, 1995 – study by Mentzer and Kahn (1995); M et al, 2006 – study by McCarthy, Davis, Golicic and Mentzer (2006); F&G, 2007 – study by Fildes and Goodwin (2007).

Data in the Table 1 reveals that preferences towards metrics have changed over the years. In the 1980s, the prevalence of the MSE/RMSE was quite clear with 34 percent – almost twice as high as of

the other two metrics. However, in the 1990s, MAPE moved in the leading position and kept it in the 2000s with over 40 percent. MAE retains the second place in all surveys. It should be noted that surveys illustrated in Table 1 were conducted using different methodologies (e.g., types of respondents, sample sizes, acceptance of multiple selections, etc.). So, the comparative results should be treated as qualitative trends rather than exact numbers.

Even most popular metrics have been scrutinized from time to time and strongly criticized or even rejected. Here are some examples.

Armstrong and Collopy (1992) stated that RMSE (arguably one of the top-used metrics) was not reliable, and was inappropriate for comparing accuracy across time series. Later, Willmott and Matsuura (2005, p. 82) found that RMSE has "disturbing characteristics" and is inappropriate for use as an error measure (Willmott, Matsuura and Robeson, 2009). The authors extended their conclusion on all square error measures (e.g., standard error). They recommended RMSE not to be reported in the literature and strongly advised in favour of using MAE. Chai and Draxler (2014) disputed these conclusions, at least partially, and presented arguments against avoiding RMSE.

Makridakis (1993) criticized the use of RAE as not meaningful for decision making.

Foss, Stensrud, Kitchenham and Myrtveit (2003) concluded that MMRE (MAPE), another very popular metric, is unreliable and may be misleding. Still, according to a number of surveys reviewed by Gneiting (2011), MAPE is the most commonly used measure for assessing forecasts in organisations.

Li (2017) asserted that correlation coefficient (R) and the coefficient of determination (R^2) should not be used as measures to assess the accuracy of predictive models for numerical data (because they are biased, insufficient or misleading).

Discussions on which metric to use are common in the literature. Usually, they are based on the premise that there could be a single "ideal" metric that beats all others in all situations. Paradoxically, a drive for having a best single metric, leads to an opposite result – the number of metrics tend to increase steeply.

Recently, new metrics are being developed and published on more regular basis (e.g., Bratu, 2013; Chen, Twycross and Garibaldi, 2017; Grigsby et al, 2018; Kim and Kim, 2016; Kyriakidis, et al, 2015; Mathai, Agarwal, Angampalli, Narayanan and Dhakshayani, 2016; Tofallis, 2015). Two approaches are commonly used to develop new metrics. First is focused on modifying existing measures to adjust them to task-specific conditions (e.g., Bratu, 2013; Grigsby et al, 2018; Mathai, et al, 2016; Monero, Pol, Abad and Blasco, 2013). The second approach is to combine the information contained in several existing measures (e.g., Kyriakidis, et al, 2015).

Still, no consensus on the "best" metric has been achieved. On the contrary, another notion is gaining popularity. Researchers express a more practical view that there is no need to strive for a single best metric. This is an unrealistic goal - "a quest for an ideal". Silver, Pyke and Thomas (2016, ch. 3) argued that "no single measure is universally best". Chai and Draxler (2014, p. 1248) clarified that "as every statistical measure condenses a large number of data into a single value, it only provides one projection of the model errors emphasizing a certain aspect of the error characteristics of the model performance." This notion is supported by Armstrong and Collopy (1992), Mahmoud (1987), Fildes and Goodwin (2007), Kyriakidis et al, (2015), etc.

There is a foundational point which needs to be mentioned considering performance metrics. Evaluation error (deviation of actual and predicted values) is a random variable. Its complete description is possible only with probability density function or moments, if they exist (e.g., Ayyub & McCuen, 2016).

Certain terminology clarifications are provided in the next three paragraphs for better understanding throughout the paper.

Some popular metrics are referred to as *scale-dependant* (Hyndman, 2006) or *dimensioned* (Willmott and Matsuura, 2005) as errors have physical dimensions and expressed in the units of the data under analysis (variable of interest), e.g., MAE, RMSE. Note that the condition to categorize a metric as dimensional is two-fold: first, it must have a dimension, and, second, the dimension must be the same as of the variable of interest. For example, if we use machine learning regression to predict cost of a medical intervention, measured in dollars, then the mean absolute error will also be found in dollars. By the same token, predicting quantities with dimensions in time, speed, distance, etc. measured in dimensional units, respectively, second, mile per hour, kilometer, etc., metrics will preserve the same units.

Two caveats need to be considered. First, certain metrics, although bear physical dimension, e.g., MSE and other squared error metrics, strictly speaking, should not be included in the dimensioned group, because their dimensions are different (changed) from the dimension of the variable of interest. For example, the cost prediction exercise mentioned above, will result in MSE measured in "squared dollars". Second, certain variables of interest have no physical dimension, i.e., dimensionless (Dimensionless Quantity, n.d.). Examples of dimensionless quantities include: GDP ratio, coefficient of determination, elasticity, etc. (List of Dimensionless Quantities, n.d.). Sometimes dimensionless quantities are given special names: percentages, degrees, decibels, radians, etc. Applying metrics to dimensionless variables of interest will provide dimensionless results. Paradoxically, applying MAE, RMSE metrics in these cases are still usually included in a dimensioned group. So the underline idea is that the metric should not be *changing* the nature (dimensional or dimensionless) of the input data.

By contrast, there is another group of metrics that do not have dimension and referred to as dimensionless (Dimensionless Quantity, n.d.) or scale-free, scaled, or scale-independent. Commonly, dimensionless metrics involve mathematical division of quantities of the same dimensional units (e.g., ratios, relative, percentage indicators), e.g., MAPE.

Part II: Known Metrics Classifications, Their Benefits, and Drawbacks

It should come as no surprise, that with a multitude of available performance metrics, research efforts are taken to organize them into categories according to common characteristics and properties for easier study, design and thoughtful application. In this review, descriptions of known metrics' classifications, their benefits and drawbacks have deliberately made rather concise. This can be explained by the fact that all prior classifications were constructed without well-established foundations, i.e., they reveal lack of explicit statements of properties that were used to group or separate certain categories of metrics.

Makridakis and Hibon (1995) proposed a classification of error metrics by two criteria: the character of measure (absolute, relative to a base or other method, relative to the size of errors) and the type of evaluation (a single method, more than one method, in comparison to some benchmark). They presented results in a table format: character of measures as rows and types of evaluation as columns. They applied the classification to a set of 14 metrics they studied and placed metric titles in the cells of intersecting criteria. It can be seen from the table that some metrics (e.g., MAPE and MdAPE) were assigned to two cells. It reveals that the classification criteria are not mutually exclusive (overlapping) which is not good for a classification. To the best of our knowledge, this was the first attempt to build a formal error metrics typology.

Hyndman (2006) suggested classifying metrics into four groups:

- scale-dependent metrics (e.g., MAE, GMAE);
- percentage-error metrics (e.g., MAPE);
- relative-error metrics (e.g., MdRAE, GMRAE);

- scale-free error metrics (e.g., MASE).

This classification is simple, intuitively clear (at least for some metrics) and has been widely used in the literature. However, in the logical sense, this classification is not perfect – it has overlappings. It appears that the groups are categorised based on whether the metric has a scale (i.e., measured in certain units) or not. Following this logic, the classification should consist of only two top-level classes: scale-dependent and scale-free. Percentage and relative metrics should be included in the scale-free metrics. Further, percentage metrics should be a subclass of the more general relative metrics (at least linguistically, although algorithmic relationship could be more complicated).

Also, it should be noted that Hyndman (2006, p. 44) includes MSE into scale-dependent group (claiming that the error is "on the same scale as the data" in the data set). This requires clarification because the MSE has a dimension of the squared scale/unit. To bring the MSE to the scale of the data we need to take a square root which results in another metric – RMSE.

Similar, but slightly different, classification was proposed by Hyndman and Koehler (2006). It acknowledged the following five groups:

- scale-dependent measures (e.g., MSE, RMSE, MAE, MdAE);
- measures based on percentage errors (e.g., MAPE, MdAPE, RMSPE, RMdSPE, sMAPE, sMdAPE);
- measures based on relative errors (e.g., MRAE, MdRAE, GMRAE);
- relative measures (e.g., RelMAE, CumRAE);
- scaled errors (e.g., MASE, RMSSE, MdASE).

This classification delineates relative metrics into measures based on relative individual errors and metrics based on combination of measures (dividing one metric by another).

Cha (2007) analyzed similarity measures as they apply to the comparison of the probability density functions. He suggested a classification which included nine groups:

- L_p Minkowski family measures (e.g., Euclidean, City block (Manhattan), Chebyshev);
- L_1 family measures (e.g., Average Manhattan – otherwise referred to as mean character distance or mean absolute error or Gower, Kulczynski distance, Soergel distance). They are based on Manhattan normalized absolute difference;
- Intersection family (e.g., Wave Hedges, Czekanowski);
- Inner product family (e.g., Kumar-Hassebrook, Dice);
- Fidelity family or Squared-chord family (e.g., fidelity, Bhattacharyya);
- Squared L_2 family (e.g., squared Euclidean, Neyman);
- Shannon's entropy family (e.g., Kullback-Leibler, Jeffreys);
- Combinations – measures utilizing multiple approaches from previous groups;
- Vicissitude measures (e.g., Vicis-Wave Hedges, Vicis symmetric).

This publication is widely cited (over 1,200 citations as of July 2018). However, the criteria of grouping metrics into categories were not explicitly stated, and there were some inconsistencies in assigning measures to the groups. For example, generalized Minkowski measure is listed as a separate measure in the Minkowski family. Some groups include distances from other groups, e.g., Family L_1 includes distances from the Intersection family, and Squared L_2 family includes distances from the Inner Product Family.

Cha's classification has been applied in several studies. Prasath et al (2017) used Cha's (2007) classification (with the exception of the Intersect family) to study 54 (fifty four) distance and similarity measures effect on the performance of K-Nearest Neighbor Classifier. Tschopp and Hernandez-Rivera (2017) used Cha's (2007) classification to study similarity and distance measures for vector-based datasets (e.g., histograms, signals, probability distribution functions). Hernández-Rivera, Cole-

man and Tschopp (2017) used Cha's (2007) classification to study similarity measures in application to X-ray diffraction patterns.

Cunningham (2009) developed a taxonomy of similarity mechanisms for case-based reasoning which includes four groups:

- Direct mechanisms (e.g., Minkowski, Manhattan, Euclidean);
- Transformation-based mechanisms (e.g., Edit Distance (Levenshtein Distance), alignment measures for biological sequences, Earth Mover Distance);
- Information theoretic measures (e.g., compression-based similarity, GenComress);
- Emergent measures arising from an in-depth analysis of the data (e.g., Random Forest, Cluster Kernels).

Jousselme and Maupin (2012) researched dissimilarity measures within the mathematical framework of evidence theory and presented a classification and general formulations for each category of measures. Their classification includes five categories/families. Four categories are the same as in Cha's classification (2007): Minkowski, Inner product, Fidelity and Information-based (Shannon). The fifth one is a Composite family based on the notion of two combined components: one that represents a measure of structural dissimilarity and the second that measures "information change relatively to orthogonal sum" (Jousselme and Maupin, 2012, p. 123).

Shcherbakov et al (2013) used a forecast error classification which is similar to Hyndman and Koehler's (2006) and included seven groups: absolute forecasting errors, measures based on percentage errors, symmetric errors, measures based on relative errors, scaled errors, relative measures and other error measures.

Weller-Fahy, Borghetti and Sodemann (2015) surveyed distance and similarity measures used within network intrusion anomaly detection. They grouped distance measures into four types:

- Power distances which are based on mathematical expressions involving raising to power (e.g., Euclidean, Manhattan, Mahalanobis, Heterogeneous distance);
- Distances on distribution laws (probability-related) (e.g., Bhattacharya coefficient, Jensen, Hellinger);
- Correlation similarities and distances (e.g., Spearman, Kendall, Pearson);
- Other similarities and distances which do not fit into the three main categories).

Some authors, without attempting to build a complete taxonomy, suggest grouping metrics by certain aspects, e.g., characteristic of error measured. Morley, Brito and Welling (2018) grouped metrics by the nature of measured statistic: accuracy (e.g., MSE, RMSE, MdAE, etc.) and bias (e.g., ME, MPE, etc.).

The review of the existing classifications revealed that their drawbacks are caused by lack of explicitly stated metrics properties that were used to group certain categories of metrics. That led to overlappings of groups and inconsistencies in assigning metrics to the categories. This study attempts to overcome revealed drawbacks of prior typologies.

FINDINGS

PERFORMANCE METRICS FRAMEWORK

Based on the analysis of the structure of numerous performance metrics presented in the literature, we propose a framework of metrics: primary metrics, extended metrics, composite metrics, and hybrid sets of metrics. Outline and examples of each category follow.

Primary metrics is arguably the most numerous category and include commonly used metrics such as MAE, MSE, sMAPE, etc. As it is shown in the next section, the structure of the primary metrics

involves three steps: calculating point distance, performing normalization and aggregating point results over a data set. Refer to the next section for detailed description and analysis. Also, these metrics are used for construction of the metrics in other categories.

Extended metrics are commonly based on the primary metrics with additional normalization. The delineation with primary metrics is that normalization is performed after aggregation. Examples include:

- Normalized Root Mean Squared Error: NRMSE_sd = RMSE/sd -normalized by the standard deviation of the actual data; or NRMSE_max-min =RMSE/(maxA – minA) - normalized by the difference between maximum and minimum actual data; or NRMSE_m = RMSE/$\bar{A}$ -normalized by the mean of actual data, also known as coefficient of variation of the RMSE (CVRMSE) (Aman, Simmhan and Prasanna, 2011; Aman, Simmhan and Prasanna, 2015).
- MAD/Mean ratio (Hoover, 2006; Kolassa and Schütz, 2007).

Composite metrics involve two or more primary metrics which are combined to produce a single result. Examples of composite metricsinclude:

- Mean Absolute Scaled Error: MASE=MAE/MAE_{ib}, where MAE_{ib} is MAE from an in-sample naïve forecast (Hyndman and Koehler, 2006).
- Relative Mean Absolute Scaled Error: RelMAE = MAE/MAE_b, where MAE_b is MAE from a benchmark method, e.g., Hyndman and Koehler (2006), and relative geometric root mean square error (RGRMSE) (Syntetos and Boylan, 2005).
- Relative Root Mean Squared Error: RelRMSE = RMSE/$RMSE_b$,where $RMSE_b$ is RMSE from a benchmark method, e.g., Chen, Twycross and Garibaldi (2017), Thomakos and Nikolopoulos (2015). Note that RelRMSE is also known as Theil's U or U2 (De Gooijer and Hyndman, 2006).

Syntetos and Boylan (2005) observed that metrics which have a term 'relative' in their title can be built by combining any methods and suggested to group them into 'accuracy measures relative to another methods'.

Vogt, Remmen, Lauster, Fuchs, and Müller (2018) tested combinations of up to six metrics in the dynamic simulation of buildings energy consumption. They recommended a composite metric calculated as a sum of four equally weighted statistical indices: the Coefficient of Variation of Root Mean Square Error (CV(RMSE)), the Normalized Mean Error (NME), the standardized contingency coefficient, and the coefficient of determination.

Hybrid sets of metrics are represented by several metrics (two or more) which are used in the same experiment with several output results. These sets are not intended to be combined in a single mathematical structure to provide a single-number output. Not any list of metrics can constitute a hybrid set. In a hybrid set, proposed metrics should be used to deliver mutually complementary properties providing better understanding of performance errors, e.g., measuring bias and accuracy. Using hybrid sets is in line with Fildes and Goodwin's (2007) advice of using multiple forecasting accuracy measures.

Kyriakidis et al (2015) developed a set of performance indices to evaluate artificial neural network models for air quality forecasting.

Another hybrid set of metrics was introduced by Morley, Brito and Welling (2018). They proposed two new metrics to be used in conjunction in radiation belt electron flux modeling and forecasting: the median symmetric accuracy and the symmetric signed percentage bias the use.

Zhang et al (2015) were searching for a set comprehensive, consistent, and robust metrics to assess performance of solar power forecasts. They recommended a suite of metrics consisting of MBE, standard deviation, skewness, kurtosis, distribution of forecast errors, Re´nyi entropy, RMSE, and OVERPer.

In our view, development of the hybrid sets of metrics should be on the top of the research agenda. Items of the agenda may include: studies on informational relationships of metrics; developing recommendations on avoiding redundancy of metrics compiled into a hybrid set; exploring ways of building minimum sets of metrics sufficiently describing error performance (e.g., Tian, Nearing, Peters-Lidard, Harrison and Tang, 2016).

PRIMARY METRICS TYPOLOGY

Analysis of multiple performance metrics used for evaluation in many fields led to identification of three (3) key components (dimensions) that determine the properties of metrics and can be used for designing typology:

- Method of determining point distance, $\mathbb{D}$.
- Method of normalization, $\mathbb{N}$.
- Method of aggregation of point distances over a data set, $\mathbb{G}$.

This approach to building a typology is usually referred to as morphological typology - a scientific method widely used in many fields, especially in linguistics, biology, astronomy, etc.

A generic formula defining a primary performance metric can be written as follows:

$$\mathbb{mm} = \mathop{\mathbb{G}^{z}}_{j=1,n}\{\mathbb{N}^{z}[\,\mathbb{D}^{z}(A_j, P_j)]\}$$

whereA_j – actual value; P_j – predicted value; n – size of the data set; z – numerical index of the method (not 'to the power of' symbol).

The meaning of the formula is in sequential determining the point distance between the actual and predicted values, normalizing it and then aggregating over a complete data set. All performance metrics explicitly contain components of determining the point distance and aggregation. Normalization component is optional, i.e., in some metrics $\mathbb{N} = 1$.

Note that to simplify notation, we are not using superscript in the individual realizations of the methods, i.e., for $z = 1$ we write $\mathbb{D}1$, not $\mathbb{D}^{1}$.

Table 2 demonstrates most common methods which will be described in the subsections below. The fact that each category has almost the same number of options (4-5) is just a coincidence. The list of methods in the typology is not intended to be comprehensive. Only most popular methods are included.

Table 2. Performance metrics typology components and implementation options

Point Distance, $\mathbb{D}$	**Normalization, $\mathbb{N}$**	**Aggregation, $\mathbb{G}$**
Error (magnitude of error): $\mathbb{D}1 = A_j - P_j$	Unitary normalization:$\mathbb{N}1 = 1$	Mean aggregation, $\mathbb{G}1$
Absolute error: $\mathbb{D}2 = \lvert A_j - P_j \rvert$	Normalization by actuals: $\mathbb{N}2 = A_j^{-c}$	Median aggregation, $\mathbb{G}2$
Squared error: $\mathbb{D}3 = (A_j - P_j)^2$	Normalization by variability of actuals: $\mathbb{N}3 = (A_j - \bar{A})^{-c}$	Geometric mean aggregation, $\mathbb{G}3$
Logarithmic quotient error: $\mathbb{D}4 = \ln(P_j/A_j)$	Normalization by the sum of actuals and predicted values: $\mathbb{N}4 = (A_j + P_j)^{-c}$	Sum aggregation, $\mathbb{G}4$
Absolute Log quotient error: $\mathbb{D}5 = \lvert\ln(P_j/A_j)\rvert$	Normalization by maximum (or minimum) value of actuals or predicted: $\mathbb{N}5 = [max(A_j, P_j)]^{-c}$	

Note: The values of the variable *c* will be explained in the next section.

Methods of determining a point distance, $\mathbb{D}$.

It should be noted that the method used to calculate point distance largely determines the overall properties of the performance metric.

In general, point distance can be calculated using any basic mathematical operation: subtraction, addition, multiplication and division (e.g., Deza and Deza, 2016). Commonly, point distances are referred to by the name of the result of the operation, respectively, difference, sum, product, quotient.

Point distances based on subtraction most commonly used in performance metrics and include: error (magnitude of error), $A_j - P_j$; absolute error, $|A_j - P_j|$; and squared error, $(A_j - P_j)^2$. They may be referred to as difference errors (e.g., Willmott et al, 1985) or just 'errors' as this type by far, the most widely used measure of error in literature.

Subtraction point distances (absolute error and squared error) correspond to the mathematical notions of the Manhattan distance (Taxicab geometry, n.d.) and the Euclidean distance (n.d.), respectively, and their generalization - Minkowski distance (n.d.). More methodological details are provided by McCune, Grace and Urban (2002).

Point distances based on division (similarly to the subtraction distances) include: magnitude of quotient error, $q_j = P_j/A_j$ (referred to as accuracy ratio by Toffalis (2015)); absolute quotient error, $|q_j|$; and squared quotient error, ${q_j}^2$. Note that division point distances are undefined when actual values are zeros.

Kitchenham, Pickard, MacDonell and Shepperd (2001) introduced quotient error (accuracy ratio) into software development effort forecasting industry, designating it variable z. Although this metric has been studied earlier, as an alternative to subtraction-type errors, in different environments. For example, Olver (1978) used it as an error for basic operations in floating-point arithmetic; also Törnqvist, Vartia and Vartia (1985) considered this metric as one of relative measures in statistics.

Most commonly, quotient error is used in the form of logarithmic quotient, i.e., $\ln(P_j/A_j)$. Although Tofallis (2015) studied squared quotient error as a loss function in prediction model selection.

Multiplication point distances are more suitable for vector represented data and binary data which are not in scope of this study. Examples can be found in inner product and fidelity groups of metrics in Cha (2007) and Prasath et al (2017): e.g., Inner Product Distance (IPD), Harmonic Mean Distance (HMD) (not to be confused with harmonic mean – aggregation procedure).

To the best of our knowledge, addition point distances were not used in the practical applications in the fields of our interest.

Properties of the commonly used point distances are outlined below.

Error (magnitude of error): $\mathbb{D}1 = A_j - P_j$

The most "natural" method of determining point distance between the actual and predicted values is subtracting one from another. The result of subtraction is a magnitude of error (or just error). Following the currently accepted notation in forecasting, we will be subtracting predicted value from the actual.

Finding the magnitude of error is a straight forward and computationally efficient method. Other methods of determining point distance use the magnitude of error for further processing.

Also, the error is measured with the same units as the data under analysis (variable of interest). It is easily interpretable. In many problems, our business objective or loss function is proportional to the difference between the actual and predicted values (not square or absolute value of this difference, as other point distances imply).

The issue with this method may arise at the aggregation phase, when the positive and negative errors will be cancelling each other. It means that even with large (but having different signs) errors the result of calculating the performance metric may yield zero demonstrating a falsely high accuracy. On another hand, this property of a magnitude of error (showing the direction of error) may convey useful information, e.g., it may be used in analysis to determine whether the forecasting method tends to overestimate or underestimate actual values, i.e., biased. This distance is used in ME, MPE, etc.

Absolute error:$\mathbb{D}2 = |A_j - P_j|$

The idea behind the absolute error is to avoid mutual cancellation of the positive and negative errors. Absolute error has only non-negative values which facilitates aggregation of point distances over the data set.

By the same token, avoiding potential of mutual cancelations has its price – skewness (bias) cannot be determined.

Absolute error preserves the same units of measurement as the data under analysis and gives all individual errors same weights (as compared to squared error). This distance is easily interpretable and when aggregated over a dataset using an arithmetic mean has a meaning of average error.

The use of absolute value might present difficulties in gradient calculation of model parameters (Chai and Draxler, 2014). This distance is used in such popular metrics as MAE, MdAE, etc.

Squared error:$\mathbb{D}3 = (A_j - P_j)^2$

Squared error follows the same idea as the absolute error – avoid negative error values and mutual cancellation of errors.

Due to the square, large errors are emphasized and have relatively greater effect on the value of performance metric (if $e > 1$). At the same time, the effect of relatively small errors ($e < 1$) will be even smaller. Sometimes this property of the squared error is referred to as penalizing extreme errors or being susceptible to outliers. Based on the application, this property may be considered positive or negative. For example, emphasizing large errors may be desirable discriminating measure in evaluating models (Chai and Draxler, 2014).

Squared error has unit measure of squared units of data. This may not be intuitive, e.g., squared dollars. This could be reversed at the aggregation phase by taking square root.

Squared error is acknowledged for its good mathematical properties. It is continuously differentiable which facilitates optimization.

Logarithmic quotient error:$\mathbb{D}4 = \ln(P_j/A_j) = \ln(P_j) - \ln(A_j)$

Logarithmic (Log) quotient error has some useful properties. The error is symmetric (to the change of actual and predicted values in the formula) and dimensionless (e.g., Tofallis, 2015; Tornqvist et al, 1985).

As an example, log quotient distance is used in Median Log Accuracy Ratio (MdLAR) or MdLQ – in author's notation (Morley, 2016; Morley, Brito and Welling, 2018).

Also, quotient distance is used (with normalization which results in non-symmetry) in the Shannon's or entropy-type metrics, e.g., Kullback-Leibler Divergence (KLD) and Jeffreys Divergence (JD) (Cha, 2007; Kullback and Leibler, 1951). Martin, Moreno, Garrido and Blanco (2015) found that the KLD-based method in the presence of contaminated noise outperformed the L2-based measure in the global localization of mobile robots experiment.

Absolute Log quotient error:$\mathbb{D}5 = |\ln(P_j/A_j)|$

The intention of taking an absolute value of the log quotient error is to ensure symmetric behaviour of the metric in a sense of possible changing the positions of the predicted and actual values in the formula without altering the result (Morley, Brito and Welling, 2018).

This distance is used in median symmetric accuracy (MdSA) which was developed to enhance certain characteristics of the MAPE (Morley, Brito and Welling, 2018).

Yu, Eder, Dennis, Chu and Schwartz (2006) proposed two metrics for evaluating air quality models using absolute log quotient error: Mean Normalized Absolute Factor Error (MNAFE) and Mean Normalized Factor Bias (MNFB).

Other point distances

Two more distance metrics have been mentioned in the literature but have not been widely used. First, a time-distance measure of accuracy designed to perform two-dimensional comparisons of time series (Granger and Jeon, 2003; Sicherl, 1994). Second, so called, mean-based measures where error is calculated as $e = \bar{A} - P_j$ for evaluating forecasts against the mean of the underlying process of intermittent demand (Prestwich, Rossi, Tarim and Hnich, 2014). These measures did not gain popularity and have not been included in the final typology.

Methods of normalization, $\mathbb{N}$.

The main idea behind normalization is to design metrics which can be used to compare multiple series having various dimensions. Most of the normalization methods involve division/multiplication of the point distance by certain parameter. Utilizing operation of mathematical division immediately leads to two properties: first, change of the dimension – often making the metric dimensionless, and second, risks of the denominator to become zero or close to zero and make operation impossible.

It should be emphasized that in our typology normalization is applied to the point distance (each individual error) prior to aggregation phase. There are some metrics with normalization or similar mathematical operations are applied to the aggregated error value. These cases are considered extended metrics.

Unitary normalization: $\mathbb{N}1 = 1$

Unitary normalization – division by one – does not require any calculations and has been included for the generalization purposes. A number of metrics employ unitary normalization, e.g., ME (MBE), MAE (MAD), MdAE, GMAE, MSE. These metrics sustain the dimension of the point distance. So, they are appropriate for analyzing single series, but not useful for comparing multiple series.

Normalization by actuals: $\mathbb{N}2 = A_j^{-c}$

Normalization by actuals involves division of the error by the actual value. For the magnitude of error and absolute error $c = 1$, and for the squared error $c = 2$. Also, for the absolute distance error, absolute actuals are used.

Normalization by actuals is used, for example, in MARE (referred to as MMRE - Mean Magnitude Relative Error – in software effort estimation field, e.g., Jørgensen, 2007).

Commonly, the results are multiplied by 100 to present the ratio as a percentage. Normalization by actuals is used in MPE, MAPE, MdAPE, RMSPE, RMdSPE – often referred to as percentage metrics.

Metrics with normalization by actuals are dimensionless allowing comparison of multiple series.

If actual values are zeros or very close to zeros, the metric cannot be used (undefined due to division by zero). An example of such scenario can be found in predicting intermittent (sporadic) demand

(Hyndman, 2006). To avoid a problem of division by zero, Tabataba et al (2017) suggest adding a small value (e.g., the lowest non-zero value of actual data) to A_j in the denominator, calling this algorithm a corrected MAPE (cMAPE).

Obvious analogy with normalization by actuals is normalization by predicted values. This method is mentioned in some papers, e.g., Tofallis (2015), Törnqvist, Vartia and Vartia (1985), but did not become popular in the literature. Fildes and Goodwin (2007) cautioned that inflating predicted values would distort this normalization type. Although the preference of forecaster practitioners towards actuals in denominator is not overwhelming: according to a survey by Green and Tashman (2009) 56% prefer actuals.

Normalization by variability of actuals: $\mathbb{N}3 = (A_j - \bar{A})^{-c}$

Normalization by variability of actuals includes division of the error by the difference between the actual value and mean value of all actuals. For the magnitude of error and absolute error $c = 1$, and for the squared error $c = 2$. Also, for the absolute distance error, absolute actuals are used.

Inclusion of the actuals mean $\bar{A}$ is intended to lower the risk of division-by-zero situations. Actuals mean is implemented in R packages (e.g., in *ML_metrics, metrics, rminer*). In general case, normalization can use an error from a benchmark method (usually naïve forecasting) (Hyndman, 2006).

Normalization by variability of actuals is used in RAE, MRAE, MdRAE, GMRAE, RSE, RRSE – often referred to as relative metrics.

Normalization by the sum of actuals and predicted values: $\mathbb{N}4 = (A_j + P_j)^{-c}$

Normalization by the sum of actual and predicted values involves division of the point distance by the sum of the actuals and predicted values. It was introduced in relation to MAPE. Initial intent behind this type of normalization was to make MAPE symmetric (Makridakis, 1993). However, later it was shown that the objective was not gained – sMAPE (symmetric MAPE) was still asymmetric (Goodwin and Lawton, 1999). At the same time, it seems reasonable to assume that the sum of the actuals and predicted values has less risk to be equal to zero. Several options of this normalization method exist (Symmetric Mean Absolute Percentage Error, n.d.). Popular ones use an average of theactuals and predicted values, i.e., $(A_j + P_j)/2$ (Green and Tashman, 2009) or use absolute actual and predicted values.

Normalization by the sum of actual and predicted values is used in sMAPE and sMdAPE – often referred to as 'symmetric' percentage metrics. Also, this normalization is used in FB and FAE (e.g., Yu et al, 2006).

Normalization by maximum (or minimum) value of actuals or predicted: $\mathbb{N}5 = [max(A_j, P_j)]^{-c}$

Normalization by the maximum (or minimum) amount of actuals and predicted values. In the known metrics, $c = 1$. It was introduced in relation to the so called Wave Hedges Distance (e.g., Cha, 2007; Prasath, Alfeilat, Lasassmeh and Hassanat, 2017). This normalization was proved useful in a comparative study of similarity metrics for compressed domain image retrieval (Hatzigiorgaki and Skodras, 2003).

Other normalization methods

Normalization by standard deviation or the difference of the actual and predicted values (as in Mean Normalized Absolute Factor Error - MNAFE) may be used.

All normalization methods described in this subsection have the form of a multiplier (denominator), so the generic formulae for a performance metric can be simplified:

$$\mathbb{mm} = \underset{j=1,n}{\mathbb{G}^{\mathbf{z}}}\{\mathbb{N}^{\mathbf{z}} \times \mathbb{D}^{\mathbf{z}}(A_j, P_j)\}$$

Although, implementation of more sophisticated methods in the future cannot be excluded.

Methods of aggregation of point distances over a data set, 𝔾.

Aggregation of point distances (in many cases after normalization) over a data set represents the final phase in the calculating primary performance metric.

Mean aggregation, 𝔾1

Note that we use the term 'mean' to refer to the 'arithmetic mean'. For any other types of means we add an attribute, e.g., geometric mean. Calculation of the arithmetic mean of the normalized point distances over a data set is the most popular aggregation method (Arithmetic Mean, n.d.). Finding arithmetic average of the observed errors is easy: it involves summing the values of point distances and dividing by the number of elements of the data set. It is also intuitively clear: the result represents an expected value of the error. The method is used, for example, in MPE, MRAE, MSE, etc. Mean aggregation is sensitive to outliers and skewed data. Refer to Other aggregation methods below for the versions of the mean aggregation intended to overcome issues with asymmetrical distributions of data and extreme values.

Median aggregation, 𝔾2

Computation of the median involves listing all point distances in an ordered form by their value (ascending or descending) and finding the number in the centre or the mean of two middle values, if the data set has even number of elements (Median, n.d.).

Opposite to other methods, median method can be called "aggregation" only conditionally: it is not based on some sort of bundling of all point distances of the data set and calculating an output value. The output of this method is one of the existing values of point distances (searched and found through a special procedure).

The median method is more resistant to outliers than the mean (Bakker and Gravemeijer, 2006). On the other hand, there is no clear and easy mathematical formula to describe the method, so theoretical considerations are a cumbersome task (although, computational algorithms present no difficulty and included in most statistical software packages).

The method is used, for example, in MdAE, MdRAE, sMdAPE, etc.

Geometric mean aggregation, 𝔾3

The geometric mean is defined as the n-th root of the product of the values of the data set (Geometric Mean, n.d.).

Geometric mean, as a median aggregation, is more robust to outliers than arithmetic mean aggregation (Fildes, 1992; Zhou, Zhou and Mathews, 1999;).

As the method includes operations of multiplication and root extraction, the downside of this method is that aggregation is undefined, if the point distances contain negative or zero-value elements.

Makridakis and Hibon (1995, p. 10) note an advantage of geometric means in interpreting model comparisons: if there are two geometric mean assessments, e.g., 10 and 12, then "the mean absolute errors of the second method are 20% higher than those of the first".

The method is used, for example, in GRMSE (Fildes, 1992; Newbold and Granger, 1974), GMRAE, GMAE, etc.

Sum aggregation, 𝔾4

The sum aggregation is just summing point distances to create a simple metric (Sum of Absolute Differences, n.d.). The method is used, for example, in RAE, SSE, RSE, SAD, etc.

Other aggregation methods

The harmonic mean is calculated as the reciprocal of the arithmetic mean of the reciprocals of the data set (Harmonic Mean, n.d.). The harmonic mean (as well as arithmetic and geometric) was known to ancient Greek mathematics since around 500 BC (Heath, 1981). Sometimes all three means referred to as Pythagorean Means (n.d.). However, this method is not as popular as the other two means.

The truncated mean (or trimmed mean) is a version of the arithmetic mean. It involves discarding some extreme data points at the high and low end before calculating arithmetic mean on the rest of the data set. This method appears to be more robust to outliers compared to a standard arithmetic mean, but could lead to a biased estimation, if underlying error distribution is not symmetric (Meyer & Venkatu, 2014; Truncated Mean, n.d.). Windsorized mean is similar to the truncated mean, except the extreme data points are not discarded but replaced by the next largest (or smallest) values (Winsorized Mean, n.d.).

Use of M-estimators is another method to deal with outliers and non-normal distributions which may contaminate arithmetic mean (M-estimator, n.d.). M-estimator is a robust estimator that weights the observations on the basis of their relative distance from the centre of the distribution. Monero et al (2013) proposed using Huber M-estimator to improve performance of the mean absolute percentage error metric. They called this metric Resistant MAPE or R-MAPE.

Similar to the median aggregation, sometimes maximum aggregation is used. It involves searching the maximum value in the point distances. This method is employed in Maximum Absolute Error (MaxAE) (Zhang et al, 2015).

VISUALIZING TYPOLOGY

The developed typology has been visualized using a table format. Metric components and their implementation options are shown on the right side (point distance), left side (aggregation) and top (normalization) of the chart. Each cell located on the intersection of three components defines an individual metric. For example, MRAE can be identified as 𝔻2ℕ3𝔾1 or GRMSEcan be identified as 𝔻3ℕ1𝔾3. Table 3 demonstrates that 40 primary metrics have been conveniently ordered and organized by their components shedding light on the properties of the metrics. Some cells of the chart are blank opening opportunities for designing new metrics.

Note that for better visualization the table is not comprehensive. It includes only most popular components. For example, MNFB metric is shown on the list in the Appendix C with mathematical definitions of metrics by not in the Table because it uses normalizer which is not very common.

DISCUSSION

The paper provided an overview of a wide range of performance metrics used in machine learning regression, forecasting and prognostics. A comparison of prior metrics classifications and their limitations was conducted. Prior typologies (e.g., Hyndman, 2006; Hyndman and Koehler, 2006) are based on a one-level ("flat") structure with 5-9 categories which made it difficult to organize multiple metrics without overlappings. Our typology suggests two levels with a detailed typology of primary metrics which allows incorporating more metrics than it was possible with prior classifications. Suggested typology has been shown to cover most of the commonly used primary metrics – total of over 40.

Also, prior typologies group together metrics with significant differences. For example, Hyndman's classification (2006) arranges together metrics based on different errors – absolute and squared – although these metrics have considerably different properties.

Finally, prior typologies operate with metrics taken as complete structures without going deeper into the metric construct. Our typology defines metrics components which determine metrics' properties.

Suggested in this paper generic formula for primary performance metrics is more comprehensive than used by Willmott & Matsuura (2005), as their definition can be applied only to metrics with mean-averaging type of error aggregation.

Table 3. Performance metrics (error measures) typology

Point Distance, $\mathbb{D}$	Normalization, $\mathbb{N}$					Aggregation, $\mathbb{G}$
	$\mathbb{N}1 = 1$ Unitary	$\mathbb{N}2 = A_j^{-c}$ By Actual Values	$\mathbb{N}3 = (A_j - \bar{A})^{-c}$ By Variability of Actual Values	$\mathbb{N}4 = (A_j + P_j)^{-c}$ By Sum of Actual and Predicted Values	$\mathbb{N}5 = [max(A_j, P_j))]^{-c}$ By Max (or Min) of Actual and Predicted Values	
Error (magnitude of error) $\mathbb{D}1 = A_j - P_j$	ME (MBE, bias)	MNB $c=1$ MPE=100MNB		FB $c=1$		$\mathbb{G}1$ Mean
						$\mathbb{G}2$ Median
						$\mathbb{G}3$ Geometric Mean
	MD					$\mathbb{G}4$ Sum
Absolute error $\mathbb{D}2 = \lvert A_j - P_j \rvert$	MAE (MAD)	MARE $c=1$ MAPE=100MARE	MRAE $c=1$	FAE $c=1$ sMAPE=100FAE		$\mathbb{G}1$ Mean
	MdAE	MdAPE $c=1$	MdRAE $c=1$	sMdAPE $c=1$		$\mathbb{G}2$ Median
	GMAE		GMRAE $c=1$			$\mathbb{G}3$ Geometric Mean
	SAD		RAE $c=1$	CM $c=1$	WHD $c=1$ max	$\mathbb{G}4$ Sum
Squared error $\mathbb{D}3 = (A_j - P_j)^2$	MSE RMSE = $\sqrt{\text{MSE}}$	MSPE $c=2$ RMSPE = $\sqrt{\text{MSPE}}$				$\mathbb{G}1$ Mean
		MdSPE $c=2$ RMdSPE = $\sqrt{\text{MdSPE}}$				$\mathbb{G}2$ Median
	GRMSE					$\mathbb{G}3$ Geometric Mean
	SSE ED = $\sqrt{\text{SSE}}$	NCSD $c=1$	RSE $c=2$ RRSE = $\sqrt{\text{RSE}}$	SquD $c=1$ DivD $c=2$	VSD $c=1$ min	$\mathbb{G}4$ Sum
Log quotient error $\mathbb{D}4 = \ln(P_j/A_j)$ $= \ln(P_j) - \ln(A_j)$	MdLAR					$\mathbb{G}1$ Mean
						$\mathbb{G}2$ Median
						$\mathbb{G}3$ Geometric Mean
		KLD $c=-1$				$\mathbb{G}4$ Sum
Absolute Log quotient error $\mathbb{D}5 = \lvert \ln(P_j/A_j) \rvert$	MNAFE					$\mathbb{G}1$ Mean
	MdSA					$\mathbb{G}2$ Median
						$\mathbb{G}3$ Geometric Mean
						$\mathbb{G}4$ Sum

The developed typology can inform metric selection process decision making by structuring performance metrics considerations (point distance, normalization and aggregation phases) and focusing on the key properties of the components chosen. For example, if the business or research need is to emphasize outliers, squared error and arithmetic mean should be used. However, if the business requirement is to isolate outliers, then selection of absolute error and geometric mean is desirable. In other words, the use of this typology turns selection of a metric from a browsing exercise over doz-

ens of metrics into a straightforward process of identifying point distance, normalization and aggregation methods that fit the purpose of the task.

The benefits of the developed typology, outlined above, are also applicable to the process of facilitating creation of new metrics. It should be noted that this study have not revealed recently conceived types of point distances, normalizers or aggregators - all of them existed for a while. Suggested visualization table can be used as a tool for creating new metrics by consciously choosing blank cells in the chart (an analogy with the Periodic Table of the chemical elements). The current structure of the visualization table potentially provides for 100 different metrics.

Assumptions and limitations

It has been shown that the typology developed in this paper can be applied to a wide variety of commonly used performance metrics. Although most known metrics can be easily classified with the typology, there are certain exceptions, as to every rule. These exceptions are usually related to metrics which have been developed for use in ad hoc circumstances (specific input data structure). For example, mean arctangent absolute percentage error (MAAPE) proposed by Kim and Kim (2016). MAAPE is a modification of MAPE which involves taking arctangent of the absolute error normalized by the actual values.

Our approach in this study is conceptual. We are not empirically comparing various metrics (e.g., Armstrong, & Collopy, 1992), but rather consider their qualitative properties.

We focus on machine learning regression with numerical data. Metrics for evaluating categorical, ordinal, binary types of data are not in scope (e.g., Choi, Cha, & Tappert, 2010).

Finally, within machine learning metrics the study considers only metrics used in regression algorithms. The tasks of classification or clustering may require different types of metrics (Deza & Deza, 2016).

Listed limitations are essential for the reader to understand what is not included in the study and shape expectations of generalizability of the findings. Also, these limitations were used for formulating directions of future research.

CONCLUSION

The importance and timeliness of the paper is determined by the increased interest of researchers and practitioners to improving evaluation results in machine learning regression, forecasting and prognostics. The paper overviewed multiple performance metrics and conducted a comparison of prior metrics classifications.

The main findings and results of the study include the following. The paper proposed metrics framework, which includes four (4) categories: primary metrics, extended metrics, composite metrics and hybrid sets of metrics. The paper identified three (3) key components (dimensions) that determine the structure and properties of primary metrics: method of determining point distance, method of normalization, method of aggregation of point distances over a data set. For each component, implementation options have been identified and their properties described. The paper proposed a new primary metrics typology designed around the key metrics components. The suggested typology has been shown to cover most of the commonly used primary metrics – total of over 40. A new generic mathematical formula for primary performance metrics has been proposed which implies sequential determining the point distance between the actual and predicted values, normalizing it and then aggregating results over a complete data set. Typology visualization chart has been designed which can be used as a tool for classifying and assessing existing, and creating new metrics.

The main contribution of this paper is in ordering knowledge of performance metrics and enhancing understanding of their structure and properties by proposing a new typology, generic primary metrics mathematical formula and a visualization chart. The practical significance of the paper is in the fact

that the presented findings can be used to facilitate teaching performance metrics to university students, expedite metrics selection process for practitioners and streamline new metrics development for academics.

Two future research opportunities can be conceived from the results of this paper. First, following the approach taken in this paper to model and analyze primary metrics, to continue conceptual research into the properties of the other metrics categories identified in this paper, namely: extended metrics, composite metrics and hybrid sets of metrics. Second, start an empirical study of the metrics, using R Studio or Azure Machine Learning Studio, to find associations between the conceptual properties of primary metrics and their "numerical" behavior in a wide spectrum of data characteristics and business or research requirements.

REFERENCES

Aman, S., Simmhan, Y., & Prasanna, V. K. (2015).Holistic measures for evaluating prediction models in smart grids. *IEEE Transactions on Knowledge and Data Engineering, 27*(2), 475-488. https://doi.org/10.1109/TKDE.2014.2327022

Aman, S., Simmhan, Y., & Prasanna, V. K. (2011). Improving energy use forecast for campus micro-grids using indirect indicators. In *IEEE 11th International Conference on Data mining Workshops (ICDMW), December 2011* (pp. 389-397). IEEE. https://doi.org/10.1109/ICDMW.2011.95

Arithmetic Mean.(n.d.). In *Wikipedia*. Retrieved July 19, 2018, from https://en.wikipedia.org/wiki/Arithmetic_mean

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting, 8*(1), 69-80. https://doi.org/10.1016/0169-2070(92)90008-W

Ayyub, B. M., & McCuen, R. H. (2016). Probability, statistics, and reliability for engineers and scientists (3rd ed.). Boca Raton: CRC Press.

Bakker, A., & Gravemeijer, K. P. (2006). An historical phenomenology of mean and median. *Educational Studies in Mathematics, 62*(2), 149-168. https://doi.org/10.1007/s10649-006-7099-8

Botchkarev, A. (2018a). *Evaluating performance of regression machine learning models using multiple error metrics in Azure Machine Learning Studio.* Retrieved from http://ssrn.com/abstract=3177507

Botchkarev, A. (2018b). *Evaluating hospital case cost prediction models using Azure Machine Learning Studio.* Retrieved from https://arxiv.org/ftp/arxiv/papers/1804/1804.01825.pdf

Bratu, M. (2013). New accuracy measures for point and interval forecasts: A case study for Romania's forecasts of inflation and unemployment rate. *Atlantic Review of Economics*, 1. Retrieved from https://www.econstor.eu/bitstream/10419/146573/1/776595040.pdf

Carbone, R., & Armstrong, J. S. (1982). Note. Evaluation of extrapolative forecasting methods: Results of a survey of academicians and practitioners. *Journal of Forecasting, 1*(2), 215-217. https://doi.org/10.1002/for.3980010207

Cha, S.H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences. 1*(4), 300-307. Retrieved from http://csis.pace.edu/ctappert/dps/d861-12/session4-p2.pdf

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247-1250. https://doi.org/10.5194/gmd-7-1247-2014

Chen, C., Twycross, J., & Garibaldi, J. M. (2017). A new accuracy measure based on bounded relative error for time series forecasting. *PloS ONE*, *12*(3), e0174202. https://doi.org/10.1371/journal.pone.0174202

Choi, S. S., Cha, S. H., & Tappert, C. C. (2010). A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics, 8*(1), 43-48. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.352.6123&rep=rep1&type=pdf

Cunningham, P. (2009). A taxonomy of similarity mechanisms for case-based reasoning. *IEEE Transactions on Knowledge and Data Engineering, 21*(11), 1532-1543. https://doi.org/10.1109/TKDE.2008.227

De Gooijer, J. G., & Hyndman, R. J. (2006).25 years of time series forecasting. *International Journal of Forecasting, 22*(3), 443-473. https://doi.org/10.1016/j.ijforecast.2006.01.001

Deza, M. M., & Deza, E. (2016). *Encyclopedia of distances* (4th ed.). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-52844-0

Dimensionless Quantity. (n.d.). In *Wikipedia*. Retrieved June 22, 2018, from https://en.wikipedia.org/wiki/Dimensionless_quantity

Euclidean Distance. (n.d.). In *Wikipedia*. Retrieved May 30, 2018, from https://en.wikipedia.org/wiki/Euclidean_distance

Fildes, R. (1992). The evaluation of extrapolative forecasting methods. *International Journal of Forecasting, 8*(1), 81-98. https://doi.org/10.1016/0169-2070(92)90009-X

Fildes, R., & Goodwin, P. (2007). Against your better judgment? How organizations can improve their use of management judgment in forecasting. *Interfaces, 37*(6), 570-576. https://doi.org/10.1287/inte.1070.0309

Foss, T., Stensrud, E., Kitchenham, B., & Myrtveit, I. (2003). A simulation study of the model evaluation criterion MMRE. *IEEE Transactions on Software Engineering, 29*(11), 985. https://doi.org/10.1109/TSE.2003.1245300

Geometric Mean. (n.d.). In *Wikipedia*. Retrieved July 19, 2018, from https://en.wikipedia.org/wiki/Geometric_mean

Gerber, A., Baskerville, R., & Van der Merwe, A. (2017). A taxonomy of classification approaches in IS research. In *Twenty-third Americas Conference on Information Systems,* Boston, 2017. Retrieved from https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1232&context=amcis2017

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association, 106*(494), 746-762. https://doi.org/10.1198/jasa.2011.r10138

Goodwin, P., & Lawton, R. (1999). On the asymmetry of the symmetric MAPE. *International Journal of Forecasting, 15*(4), 405-408. https://doi.org/10.1016/S0169-2070(99)00007-2

Granger, C. W. J., & Jeon, Y. (2003). A time–distance criterion for evaluating forecasting models. *International Journal of Forecasting, 19*, 199 – 215. https://doi.org/10.1016/S0169-2070(02)00030-4

Green, K., & Tashman, L. (2009). Percentage error: What denominator? *Foresight: The International Journal of Applied Forecasting*, (12), 36-40.

Grigsby, M. R., Di, J., Leroux, A., Zipunnikov, V., Xiao, L., Crainiceanu, C., & Checkley, W. (2018). Novel metrics for growth model selection. *Emerging Themes in Epidemiology, 15*(4). https://doi.org/10.1186/s12982-018-0072-z

Harmonic Mean.(n.d.). In *Wikipedia.* Retrieved July 19, 2018, from https://en.wikipedia.org/wiki/Harmonic_mean

Hatzigiorgaki, M., & Skodras, A. N. (2003, June). Compressed domain image retrieval: A comparative study of similarity metrics. In *Proceedings of Visual Communications and Image Processing* 2003 (Vol. 5150, pp. 439-449). International Society for Optics and Photonics. https://doi.org/10.1117/12.507669

Heath, T. H. (1981). *A history of Greek mathematics*. Dover, New York.

Hernández-Rivera, E., Coleman, S. P., & Tschopp, M. A. (2017). Using Similarity Metrics to Quantify Differences in High-Throughput Data Sets: Application to X-ray Diffraction Patterns. *ACS Combinatorial Science, 19*(1), 25-36. https://doi.org/10.1021/acscombsci.6b00142

Hoover, J. (2006). Measuring forecast accuracy: Omissions in today's forecasting engines and demand-planning software. *Foresight: The International Journal of Applied Forecasting, 4*(4), 32-35.

Hyndman, R. J. (2006). Another look at forecast-accuracy metrics for intermittent demand. *Foresight: The International Journal of Applied Forecasting, 4*(4), 43-46.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Jørgensen, M. (2007, December). A critique of how we measure and interpret the accuracy of software development effort estimation. In *First International Workshop on Software Productivity Analysis and Cost Estimation*. Information Processing Society of Japan, Nagoya. Retrieved from https://pdfs.semanticscholar.org/f347/d4de8a1decfcea2602c33254dbf4b5bd366d.pdf#page=25

Jousselme, A. L., & Maupin, P. (2012). Distances in evidence theory: Comprehensive survey and generalizations. *International Journal of Approximate Reasoning, 53*(2), 118-145. https://doi.org/10.1016/j.ijar.2011.07.006

Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting, 32*(3), 669-679. https://doi.org/10.1016/j.ijforecast.2015.12.003

Kitchenham, B. A., Pickard, L. M., MacDonell, S. G., & Shepperd, M. J. (2001). What accuracy statistics really measure? *IEE Proceedings-Software, 148*(3), 81-85. https://doi.org/10.1049/ip-sen:20010506

Kolassa, S., & Schütz, W. (2007). Advantages of the MAD/MEAN ratio over the MAPE. *Foresight: The International Journal of Applied Forecasting*, (6), 40-43.

Kullback, S., & Leibler, R. A. (1951).On information and sufficiency. *The Annals of Mathematical Statistics, 22*(1), 79-86. https://doi.org/10.1214/aoms/1177729694

Kyriakidis, I., Kukkonen, J., Karatzas, K., Papadourakis, G., & Ware, A. (2015). *New statistical indices for evaluating model forecasting performance*. Skiathos Island, Greece. Retrieved from http://iranarze.ir/wp-content/uploads/2017/12/53-English-IranArze.pdf

Li, J. (2017). Assessing the accuracy of predictive models for numerical data: Not r nor r2, why not? Then what? *PloS One, 12*(8), e0183250. Retrieved from https://doi.org/10.1371/journal.pone.0183250

List of Dimensionless Quantities. (n.d.). In *Wikipedia*. Retrieved June 22, 2018, from https://en.wikipedia.org/wiki/List_of_dimensionless_quantities

M-estimator. (n.d.). In *Wikipedia*. Retrieved June 22, 2018, from https://en.wikipedia.org/wiki/M-estimator

Mahmoud, E. (1987). The Evaluation of Forecasts. in *The Handbook of Forecasting: A Manager's Guide*, Makridakis, S. and Wheelwright, S. C., eds. New York: John Wiley & Sons.

Makridakis, S. (1993). Accuracy measures: Theoretical and practical concerns. *International Journal of Forecasting*, *9*(4), 527-529. https://doi.org/10.1016/0169-2070(93)90079-3

Makridakis, S., & Hibon, M. (1995). *Evaluating accuracy (or error) measures*. INSEAD Working Paper Series, Fontainebleau, France.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PloS One, 13*(3), e0194889. https://doi.org/10.1371/journal.pone.0194889

Martín, F., Moreno, L., Garrido, S., & Blanco, D. (2015). Kullback-Leibler divergence-based differential evolution Markov Chain filter for global localization of mobile robots. *Sensors, 15*(9), 23431-23458. https://doi.org/10.3390/s150923431

Mathai, A. V., Agarwal, A., Angampalli, V., Narayanan, S., & Dhakshayani, E. (2016). Development of new methods for measuring forecast error. *International Journal of Logistics Systems and Management, 24*(2), 213-225. https://doi.org/10.1504/IJLSM.2016.076472

McCarthy, T. M., Davis, D. F., Golicic, S. L., & Mentzer, J. T. (2006). The evolution of sales forecasting management: A 20-year longitudinal study of forecasting practices. *Journal of Forecasting, 25*(5), 303-324. https://doi.org/10.1002/for.989

McCune, B., Grace, J. B., & Urban, D. L. (2002). *Analysis of ecological communities (Vol. 28). Chapter 6*. Gleneden Beach, OR: MjM software design. Retrieved from https://www.umass.edu/landeco/teaching/multivariate/readings/McCune.and.Grace.2002.chapter6.pdf

Median. (n.d.). In *Wikipedia*. Retrieved July 19, 2018, from https://en.wikipedia.org/wiki/Median

Mentzer, J. T., & Kahn, K. B. (1995). Forecasting technique familiarity, satisfaction, usage, and application. *Journal of Forecasting, 14*(5), 465-476. https://doi.org/10.1002/for.3980140506

Meyer, B., & Venkatu, G. (2014). Trimmed-mean inflation statistics: Just hit the one in the middle. *Federal Reserve Bank of Cleveland Working Papers*, (WP 12-17R).

Minkowski Distance.(n.d.). In *Wikipedia*. Retrieved May 30, 2018, from https://en.wikipedia.org/wiki/Minkowski_distance

Moreno, J. J. M., Pol, A. P., Abad, A. S., & Blasco, B. C. (2013). Using the R-MAPE index as a resistant measure of forecast accuracy. *Psicothema, 25*(4), 500-506. https://doi.org/10.7334/psicothema2013.23

Morley, S. K. (2016). *Alternatives to accuracy and bias metrics based on percentage errors for radiation belt modeling applications*. Los Alamos National Laboratory report, LA-UR-16-24592. https://doi.org/10.2172/1260362

Morley, S. K., Brito, T. V., & Welling, D. T. (2018). Measures of model performance based on the log accuracy ratio. *Space Weather, 16*(1), 69-88. https://doi.org/10.1002/2017SW001669

Newbold, P., & Granger, C. W. (1974). Experience with forecasting univariate time series and the combination of forecasts.*Journal of the Royal Statistical Society*. Series A (General), *137*(2), 131-165. https://doi.org/10.2307/2344546

Olver, F. W. J. (1978). A new approach to error arithmetic. *SIAM Journal on Numerical Analysis, 15*(2), 368-393. https://doi.org/10.1137/0715024

Parrochia, D. (n.d.). *The internet encyclopedia of philosophy* (Classification), ISSN 2161-0002, Retrieved on September 6, 2018, from https://www.iep.utm.edu/classifi

Prasath, V. B., Alfeilat, H. A. A., Lasassmeh, O., & Hassanat, A. (2017). *Distance and similarity measures effect on the performance of k-nearest neighbor classifier-A review*. Retrieved from https://arxiv.org/pdf/1708.04321.pdf

Prestwich, S., Rossi, R., Tarim, S. A., & Hnich, B. (2014). Mean-based error measures for intermittent demand forecasting. *International Journal of Production Research, 52*(22), 6782-6791. https://doi.org/10.1080/00207543.2014.917771

Pythagorean Means. (n.d.). In *Wikipedia*. Retrieved July 19, 2018, from https://en.wikipedia.org/wiki/Pythagorean_means

Saxena, A., Celaya, J., Balaban, E., Goebel, K., Saha, B., Saha, S., & Schwabacher, M. (2008). Metrics for evaluating performance of prognostic techniques. In *International Conference on Prognostics and Health Management, 2008 (PHM October, 2008)*, (pp. 1-17). IEEE. https://doi.org/10.1109/PHM.2008.4711436

Shcherbakov, M. V., Brebels, A., Shcherbakova, N. L., Tyukov, A. P., Janovsky, T. A., & Kamaev, V. A. E. (2013). A survey of forecast error measures. *World Applied Sciences Journal, (Information Technologies in Modern Industry, Education & Society) 24*, 171-176.

Sicherl, P. (1994). Time distance as an additional measure of discrepancy between actual and estimated values in time series models. In *International Symposium on Economic Modelling. The World Bank, Washington DC*.

Silver, E. A., Pyke, D. F., & Thomas, D. J. (2016). *Inventory and production management in supply chains*. CRC Press. https://doi.org/10.1201/9781315374406

Sum of Absolute Differences. (n.d.). In *Wikipedia*. Retrieved July 19, 2018, from https://en.wikipedia.org/wiki/Sum_of_absolute_differences

Symmetric Mean Absolute Percentage Error. (n.d.). In *Wikipedia*. Retrieved July 4, 2018, from https://en.wikipedia.org/wiki/Symmetric_mean_absolute_percentage_error

Syntetos, A. A., & Boylan, J. E. (2005). The accuracy of intermittent demand estimates. *International Journal of Forecasting, 21*(2), 303-314. https://doi.org/10.1016/j.ijforecast.2004.10.001

Tabataba, F. S., Chakraborty, P., Ramakrishnan, N., Venkatramanan, S., Chen, J., Lewis, B., & Marathe, M. (2017).A framework for evaluating epidemic forecasts. *BMC infectious diseases, 17*(1), 345. https://doi.org/10.1186/s12879-017-2365-1

Taxicab Geometry.(n.d.). In *Wikipedia*. Retrieved May 30, 2018, from https://en.wikipedia.org/wiki/Taxicab_geometry

Tian, Y., Nearing, G. S., Peters-Lidard, C. D., Harrison, K. W., & Tang, L. (2016). Performance metrics, error modeling, and uncertainty quantification. *Monthly Weather Review, 144*(2), 607-613. https://doi.org/10.1175/MWR-D-15-0087.1

Tofallis, C. (2015). A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society*, *66*(8), 1352-1362. https://doi.org/10.1057/jors.2014.103

Thomakos, D. D., & Nikolopoulos, K. (2015). Forecasting multivariate time series with the theta method. *Journal of Forecasting, 34*(3), 220-229. https://doi.org/10.1002/for.2334

Törnqvist, L., Vartia, P., & Vartia, Y. O. (1985). How should relative changes be measured? *The American Statistician*, *39*(1), 43-46.

Tschopp, M. A., & Hernandez-Rivera, E. (2017). *Quantifying similarity and distance measures for vector-based Datasets: Histograms, signals, and probability distribution functions* (No. ARL-TN-0810). US Army Research Laboratory Aberdeen Proving Ground United States. Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/1026967.pdf

Truncated Mean.(n.d.). In *Wikipedia*. Retrieved July 19, 2018, from https://en.wikipedia.org/wiki/Truncated_mean

Vogt, M., Remmen, P., Lauster, M., Fuchs, M., & Müller, D. (2018). Selecting statistical indices for calibrating building energy models. *Building and Environment. 144*, 94-107. https://doi.org/10.1016/j.buildenv.2018.07.052

Weller-Fahy, D. J., Borghetti, B. J., & Sodemann, A. A. (2015). A survey of distance and similarity measures used within network intrusion anomaly detection. *IEEE Communications Surveys & Tutorials, 17*(1), 70-91. https://doi.org/10.1109/COMST.2014.2336610

Willmott, C. J., Ackleson, S. G., Davis, R. E., Feddema, J. J., Klink, K. M., Legates, D. R., O'Donnell, J., & Rowe, C. M. (1985). Statistics for the evaluation and comparison of models. *Journal of Geophysical Research: Oceans, 90*(C5), 8995-9005. https://doi.org/10.1029/JC090iC05p08995

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*, *30*(1), 79-82. https://doi.org/10.3354/cr030079

Willmott, C. J., Matsuura, K., & Robeson, S. M. (2009). Ambiguities inherent in sums-of-squares-based error statistics. *Atmospheric Environment*, *43*(3), 749-752. https://doi.org/10.1016/j.atmosenv.2008.10.005

Winsorized Mean.(n.d.). In *Wikipedia*. Retrieved July 19, 2018, from https://en.wikipedia.org/wiki/Winsorized_mean

Yu, S., Eder, B., Dennis, R., Chu, S. H., & Schwartz, S. E. (2006). New unbiased symmetric metrics for evaluation of air quality models. *Atmospheric Science Letters, 7*(1), 26-34. https://doi.org/10.1002/asl.125

Zhang, J., Florita, A., Hodge, B. M., Lu, S., Hamann, H. F., Banunarayanan, V., & Brockway, A. M. (2015). A suite of metrics for assessing the performance of solar power forecasting. *Solar Energy, 111*, 157-175. https://doi.org/10.1016/j.solener.2014.10.016

Zhou, Q. H., Zhou, Q. N., & Mathews, J. D. (1999). Arithmetic average, geometric average, and ranking: Application to incoherent scatter radar data processing. *Radio Science, 34*(5), 1227-1237. https://doi.org/10.1029/1999RS900062

ACKNOWLEDGMENT

The views, opinions and conclusions expressed in this paper are those of the author alone and do not necessarily represent the views of his current or former employer(s) or organizations he is affiliated with.

APPENDICES

APPENDIX A: LIST OF METRICS ABBREVIATIONS

Metric Abbreviation	Metric Name
CM	Canberra Metric
CoD	Coefficient of Determination
CVRMSE	Coefficient of variation of the RMSE
DivD	Divergence Distance
ED	Euclidean Distance (L2-norm)
FAE	Fractional absolute error
FB	Fractional Bias
GMAE	Geometric Mean Absolute Error
GMRAE	Geometric Mean Relative Absolute Error
GRMSE	Geometric Root Mean Squared Error
HMD	Harmonic Mean Distance (not to be confused with harmonic mean – aggregation procedure)
IPD	Inner Product Distance
JD	Jeffreys Divergence
KLD	Kullback-Leibler Divergence
LMR	Log Mean Squared Error Ratio
MAAPE	Mean Arctangent Absolute Percentage Error
MAD	Mean Absolute Deviation
MAE	Mean Absolute Error
MAGE	Mean Absolute Gross Error
MAPE	Mean Absolute Percentage Error
MARE	Mean Absolute Relative Error
MASE	Mean Absolute Scaled Error
MaxAE	Maximum Absolute Error
MBE	Mean Bias Error
MCD	Mean Character Difference
MD	Manhattan Distance
MdAE	Median Absolute Error
MdAPE	Median Absolute Percentage Error
MdASE	Median Absolute Scaled Error
MdLAR	Median Log Accuracy Ratio
MdRAE	Median Relative Absolute Error
MdSA	Median Symmetric Accuracy
MdSPE	Median Square Percentage Error
ME	Mean Error
MMRE	Mean Magnitude Relative Error
MNAFE	Mean Normalized Absolute Factor Error
MNB	Mean Normalized Bias

MNFB	Mean Normalized Factor Bias
MPE	Mean Percentage Error
MRAE	Mean Relative Absolute Error
MSE	Mean Squared Error
MSPE	Mean Square Percentage Error
NCSD	Neyman Chi-Square Distance
NMSE	Normalized Mean Squared Error (normalized by variance)
NRMSE_m	Normalized Root Mean Squared Error (normalized by the mean of actual data)
NRMSE_mm	Normalized Root Mean Squared Error (normalized by the difference between maximum and minimum actual data)
NRMSE_sd	Normalized Root Mean Squared Error (normalized by the standard deviation of the actual data)
RAE	Relative Absolute Error
RelRMSE	Relative Root Mean Square Error
RMAE	Relative Mean Absolute Error
RMdSPE	Root Median Square Percentage Error
RMSE	Root Mean Squared Error
RMSPE	Root Mean Square Percentage Error
RMSSE	Root Mean Squared Scaled Error
RRSE	Root Relative Squared Error
RSE	Relative Squared Error
SAD	Sum of absolute differences
sMAPE	Symmetric Mean Absolute Percentage Error
SMdAPE	Symmetric Median Absolute Percentage Error
SquD	Squared Chi-square Distance
SSE	Sum of Squared Error (Squared Euclidean)
VSD	Vicis Symmetric Distance
WHD	Wave Hedges Distance

APPENDIX B: METRICS MATHEMATICAL DEFINITIONS

Note 1. Legend:A_j – actual values; $\bar{A}$ – the mean of the actual values; P_j – predicted values; $e_j = A_j - P_j$ – error; n – size of the data set

Note 2. Metrics are listed according to the categories they belong to, i.e., primary, extended, composite, hybrid sets; and within categories – by type of error.

Metric Abbreviation	**Metric Name** (alternative names are given in brackets)	**Metric Formula**
PRIMARY METRICS		
	Error (magnitude of error): $\mathbb{D}1 = A_j - P_j = e_j$	
ME	Mean Error (Mean Bias Error)	$ME = \frac{1}{n}\sum_{j=1}^{n} e_j$
MNB	Mean Normalized Bias	$MNB = \frac{1}{n}\sum_{j}^{n} \frac{e_j}{A_j}$
MPE	Mean Percentage Error	$MPE = \frac{100}{n}\sum_{j}^{n} \frac{e_j}{A_j}$
FB	Fractional Bias	$FB = \frac{1}{n}\sum_{j}^{n} \frac{2 * e_j}{A_j + P_j}$
MD	Manhattan Distance (City Block, L_1-norm, Taxicab norm)	$MD = \sum_{j=1}^{n} e_j$
	Absolute error: $\mathbb{D}2 = \lvert A_j - P_j \rvert = \lvert e_j \rvert$	
MAE	Mean Absolute Error (Mean Absolute Deviation – MAD; Mean Absolute Gross error; Mean Character Difference – MCD; Average Manhattan; Gower)	$MAE = \frac{1}{n}\sum_{j=1}^{n} \lvert e_j \rvert$
MdAE	Median Absolute Error	$MdAE = \underset{j=1,n}{Md}(\lvert e_j \rvert)$
MaxAE	Maximum Absolute Error	$MaxAE = \underset{j=1,n}{max}(\lvert e_j \rvert)$
MARE	Mean Absolute Relative Error (Mean Magnitude Relative Error – MMRE)	$MARE = \frac{1}{n}\sum_{j}^{n} \frac{\lvert e_j \rvert}{\lvert A_j \rvert}$
MAPE	Mean Absolute Percentage Error	$MAPE = \frac{100}{n}\sum_{j}^{n} \frac{\lvert e_j \rvert}{\lvert A_j \rvert}$
MdAPE	Median Absolute Percentage Error	$MdAPE = 100 * \underset{j=1,n}{Md}(\frac{\lvert e_j \rvert}{\lvert A_j \rvert})$

RAE	Relative Absolute Error	$RAE = \sum_{j=1}^{n} \frac{\lvert e_j \rvert}{\lvert A_j - \bar{A} \rvert}$
MRAE	Mean Relative Absolute Error	$MRAE = \frac{1}{n} \sum_{j=1}^{n} \frac{\lvert e_j \rvert}{\lvert A_j - \bar{A} \rvert}$
GMAE	Geometric Mean Absolute Error	$GMAE = \sqrt[n]{\prod_{j=1}^{n} \lvert e_j \rvert}$
SAD	Sum of Absolute Differences	$SAD = \sum_{j=1}^{n} \lvert e_j \rvert$
GMRAE	Geometric Mean Relative Absolute Error	$GMRAE = exp\left(\frac{1}{n} \sum_{j=1}^{n} \ln\left(\frac{\lvert e_j \rvert}{\lvert A_j - \bar{A} \rvert}\right)\right)$ or $= \sqrt[n]{\prod_{j=1}^{n} \left(\frac{\lvert e_j \rvert}{\lvert A_j - \bar{A} \rvert}\right)}$
MdRAE	Median Relative Absolute Error	$MdRAE = \underset{j=1,n}{Md} \left(\frac{\lvert e_j \rvert}{\lvert A_j - \bar{A} \rvert}\right)$
WHD	Wave Hedges Distance	$WHD = \sum_{j=1}^{n} \frac{\lvert e_j \rvert}{max(A_j, P_j)}$
FAE	Fractional absolute error	$FAE = \frac{1}{n} \sum_{j}^{n} \frac{2 * \lvert e_j \rvert}{\lvert A_j \rvert + \lvert P_j \rvert}$
sMAPE	Symmetric Mean Absolute Percentage Error	$sMAPE = \frac{100}{n} \sum_{j}^{n} \frac{2 * \lvert e_j \rvert}{\lvert A_j \rvert + \lvert P_j \rvert}$
SMdAPE	Symmetric Median Absolute Percentage Error	$SMdAPE = 100 * \underset{j=1,n}{Md} \left(\frac{2 * \lvert e_j \rvert}{\lvert A_j \rvert + \lvert P_j \rvert}\right)$
CM	Canberra Metric	$CM = \sum_{j}^{n} \frac{\lvert e_j \rvert}{A_j + P_j}$

	Squared error: $\mathbb{D}3 = (A_j - P_j)^2 = e_j^2$	
MSE	Mean Squared Error	$MSE = \frac{1}{n}\sum_{j=1}^{n} e_j^2$
RMSE	Root Mean Squared Error (Average Distance)	$RMSE = \sqrt{\frac{\sum_{j=1}^{n} e_j^2}{n}}$ or $RMSE = \sqrt{MSE}$
SSE	Sum of Squared Error (Squared Euclidean)	$SSE = \sum_{j=1}^{n} e_j^2$
ED	Euclidean Distance (L_2-norm)	$ED = \sqrt{\sum_{j=1}^{n} e_j^2}$ or $ED = \sqrt{SSE}$
VSD	Vicis Symmetric Distance	$VSD = \sum_{j=1}^{n} \frac{e_j^2}{min(A_j, P_j)}$
NCSD	Neyman Chi-Square Distance	$NCSD = \sum_{j=1}^{n} \frac{e_j^2}{A_j}$
SquD	Squared Chi-square Distance	$SquD = \sum_{j=1}^{n} \frac{e_j^2}{A_j + P_j}$
DivD	Divergence Distance	$DivD = 2\sum_{j=1}^{n} \frac{e_j^2}{(A_j + P_j)^2}$
RSE	Relative Squared Error	$RSE = \sum_{j=1}^{n} \frac{e_j^2}{(A_j - \bar{A})^2}$
RRSE	Root Relative Squared Error	$RRSE = \sqrt{\sum_{j=1}^{n} \frac{e_j^2}{(A_j - \bar{A})^2}}$
GRMSE	Geometric Root Mean Squared Error	$GRMSE = \sqrt[2n]{\prod_{j=1}^{n} e_j^2}$
MSPE	Mean Square Percentage Error	$MSPE = \frac{100}{n}\sum_{j}^{n} \left(\frac{\lvert e_j \rvert}{\lvert A_j \rvert}\right)^2$
MdSPE	Median Square Percentage Error	$MdSPE = 100 * \underset{j=1,n}{Md} \left(\frac{\lvert e_j \rvert}{\lvert A_j \rvert}\right)^2$
RMSPE	Root Mean Square Percentage Error	$RMSPE = \sqrt{\frac{100}{n}\sum_{j}^{n} \left(\frac{\lvert e_j \rvert}{\lvert A_j \rvert}\right)^2}$

RMdSPE	Root Median Square Percentage Error	$RMdSPE = \sqrt{100 * \underset{j=1,n}{Md}\left(\frac{\lvert e_j \rvert}{\lvert A_j \rvert}\right)^2}$
	Logarithmic quotient error: $\mathbb{D}4 = \ln(P_j/A_j) = \ln(P_j) - \ln(A_j)$	
MdLAR	Median Log Accuracy Ratio	$MdLAR = \underset{j=1,n}{Md}(\ln(P_j/A_j))$
KLD	Kullback-Leibler Divergence	$KLD = \sum_{j=1}^{n} P_j \ln(P_j/A_j)$
JD	Jeffreys Divergence	$JD = \sum_{j=1}^{n} (P_j - A_j)\ln(P_j/A_j)$
	Absolute Log quotient error: $\mathbb{D}5 = \lvert\ln(P_j/A_j)\rvert$	
MNAFE	Mean Normalized Absolute Factor Error	$MNAFE = \frac{1}{n}\sum_{j=1}^{n} \lvert\exp(\lvert\ln(\frac{P_j}{A_j})\rvert - 1\rvert$
MNFB	Mean Normalized Factor Bias	$MNFB = \frac{1}{n}\sum_{j=1}^{n} \frac{P_j - A_j}{\lvert P_j - A_j \rvert}[\exp(\lvert\ln(\frac{P_j}{A_j})\rvert - 1]$
MdSA	Median Symmetric Accuracy	$MdSA = 100(\exp(\underset{j=1,n}{Md}(\lvert\ln(P_j/A_j)\rvert))$-1)
EXTENDED METRICS		
NRMSE_m	Normalized Root Mean Squared Error (normalized by the mean of actual data) (CVRMSE - coefficient of variation of the RMSE)	$NRMSE\_m = \frac{RMSE}{\bar{A}}$
NRMSE_sd	Normalized Root Mean Squared Error (normalized by the standard deviation of the actual data)	$NRMSE\_sd = \frac{RMSE}{sd}$
NRMSE_mm	Normalized Root Mean Squared Error (normalized by the difference between maximum and minimum actual data)	$NRMSE\_mm = \frac{RMSE}{maxA - minA}$
NMSE	Normalized Mean Squared Error (normalized by variance)	$NMSE = \frac{MSE}{\sigma^2}$

COMPOSITE METRICS		
RMAE	Relative Mean Absolute Error	$RMAE=MAE/MAE\ in-sample$
RelRMSE	Relative Root Mean Square Error	$RelRMSE=$RMSE$/RMSE\ in-sample$
LMR	Log Mean Squared Error Ratio	$LMR=\log($RMSE$/RMSE\ in-sample)$
CoD	Coefficient of Determination	$$CoD = 1 - \frac{\sum_{j=1}^{n}(P_j - A_j)^2}{\sum_{j=1}^{n}(A_j - \bar{A})^2}$$
MASE	Mean Absolute Scaled Error	$MASE=MAE/MAE\ in-sample$, *naïve* $MASE=MAE/Q$, *where* $$Q = \frac{1}{n-1}\sum_{j=2}^{n}\lvert A_j - A_{j-1}\rvert$$

APPENDIX C: PERFORMANCE METRICS ALTERNATIVE MATHEMATICAL DEFINITIONS

Metric Abbreviation	**Metric Name** (alternative names are given in brackets)	**Metric Formula**
RAE	Relative Absolute Error	Option 1 $$RAE = \sum_{j=1}^{n} \frac{\lvert e_j \rvert}{\lvert A_j - \bar{A} \rvert}$$ Option 2 $$RAE = \frac{\sum_{j=1}^{n} \lvert e_j \rvert}{\sum_{j=1}^{n} \lvert A_j - \bar{A} \rvert}$$
MRAE	Mean Relative Absolute Error	Option 1 $$MRAE = \frac{1}{n} \sum_{j=1}^{n} \frac{\lvert e_j \rvert}{\lvert A_j - \bar{A} \rvert}$$ Option 2 $$MRAE = \frac{\sum_{j=1}^{n} \lvert e_j \rvert}{n \sum_{j=1}^{n} \lvert A_j - \bar{A} \rvert}$$
RSE	Relative Squared Error	Option 1 $$RSE = \sum_{j=1}^{n} \frac{e_j^2}{(A_j - \bar{A})^2}$$ Option 2 $$RSE = \frac{\sum_{j=1}^{n} e_j^2}{\sum_{j=1}^{n} (A_j - \bar{A})^2}$$
RRSE	Root Relative Squared Error	Option 1 $$RRSE = \sqrt{\sum_{j=1}^{n} \frac{e_j^2}{(A_j - \bar{A})^2}}$$ Option 2 $$RRSE = \sqrt{\frac{\sum_{j=1}^{n} e_j^2}{\sum_{j=1}^{n} (A_j - \bar{A})^2}}$$

BIOGRAPHY

Alexei Botchkarev is the Principal of GS Research & Consulting (www.gsrc.ca) and an Adjunct Professor with the Computer Science Department at Ryerson University. He holds B.Eng. five-year degree from the Kiev Aviation Engineering Academy, Ukraine (1975) and Ph.D. from the aerospace R&D institute, Russia (1985). Alexei is a public service practitioner, consultant, and researcher with contributions to simulation, implementation, and evaluation of complex systems in information management and aerospace. Results of his research are published in more than 80 journal papers, professional magazine articles, technical reports and chapters in three peer-reviewed books.