RExQUAL Stability Analysis

A. R. Troncoso-García, *Pablo de Olavide University, Spain,* M. Martínez-Ballesteros, *University of Seville, Spain,* F. Martínez-Álvarez, *Pablo de Olavide University, Spain,* A. Troncoso, *Pablo de Olavide University, Spain*

I. Introduction

The RExQUAL metric emerges as an innovative, model-agnostic solution for evaluating and comparing the quality of explanations provided by attribution-based explainable artificial intelligence techniques. Its development addresses the need for quantitative tools to analyze the effectiveness of explainability methods in prediction contexts. RExQUAL combines local and global explanations through an approach that integrates attribution-based feature selection with a broader analysis of the entire dataset. At its core, it utilizes a subset of the attribute rankings highlighted by a model-agnostic explainability method applied to forecasting tasks. Based on these key attributes, association rules are generated and evaluated using global metrics such as support and confidence to assess their quality. This analysis focuses on a critical aspect of RExQUAL: its stability. Assessing the stability of the metric involves examining its behavior under varying conditions and configurations to ensure its robustness and reliability as an evaluation tool. This approach not only strengthens the credibility of RExQUAL as a quantitative evaluation framework but also provides a reliable basis for its practical implementation.

In the context of machine learning, stability refers to a model's ability to maintain reliable performance and accurate predictions when subjected to small changes or perturbations in the input data. This property is essential for ensuring that the model or metric remains robust and predictable, even in the presence of minor noise, measurement errors, or variations commonly encountered in real-world scenarios [1]. Stability thus serves as a critical measure of reliability, highlighting a model's capacity to generalize effectively under diverse conditions. Stability is particularly important in applications where minor inconsistencies in input data are inevitable, as unstable models may lead to significant deviations in predictions, potentially causing erroneous decisions or actions. By analyzing and quantifying stability, researchers and practitioners can identify the limits of a model's resilience and make informed adjustments to improve its reliability in practical deployments.

Let $\phi(x)$ represent the prediction generated by the model for a given input instance x. Stability in this context implies that the model produces outputs that are resilient to small perturbations or changes in the input. Specifically, the system is considered stable if, for any input instance x and its perturbed version $x + \Delta x$, the absolute difference between the model´s prediction for the original input $\phi(x)$ and the prediction for the perturbed input $\phi(x + \Delta x)$ does not exceed a predefined threshold ϵ.

Mathematically, this condition is expressed as:

$$|\phi(x + \Delta x) - \phi(x)| \leq \epsilon,$$

where:

- $\phi(x)$ is the output of the model or system for the input x.
- Δx is the perturbation applied to the input x.
- ϵ is a positive constant value that defines the maximum allowable deviation for the system to be regarded as stable. This threshold ensures that the model's predictions remain within an acceptable range, even when subjected to variations in the input.

To ensure that this definition applies to controlled perturbations, the system may also impose restrictions on Δx, such as:

$$\|\Delta x\| \leq \delta,$$

where:

- $\|\Delta x\|$ represents the norm (or magnitude) of the perturbation.
- δ is a positive value defining the maximum allowable perturbation.

This ensures that the stability analysis is conducted within a controlled range of perturbations.

II. Input data

To thoroughly evaluate the stability of the explanations provided by the RExQUAL evaluation index, experiments were conducted using the electricity demand dataset.

The electric demand dataset is a univariate time series representing the demand for electrical energy in Spain, with measurements expressed in kilowatt-hours (kWh) [2]. The dataset covers a period of nine years and six months, from January 1st, 2007, to June 21st, 2016, with data recorded at 10-minute intervals. The dataset consists of 10161 instances, each corresponding to a specific time stamp, with 192 time steps representing one day and eight hours of recorded data. The input for the model is a sliding window of the past 168 values, which corresponds to a time window of one day and four hours, used to predict the demand for the subsequent 24 time steps (a 4-hour prediction horizon).

The histogram of the first column of the input is presented in Figure 1. It provides a visual representation of the distribution of electric demand values. It shows how frequently different demand levels occur within the dataset, with the x-axis representing the range of values in kWh and the y-axis indicating the frequency of each value range. The demand values typically fall within a range of approximately 15,000 to 45,000 kWh, with certain periods exhibiting higher or lower frequencies based on consumption patterns.

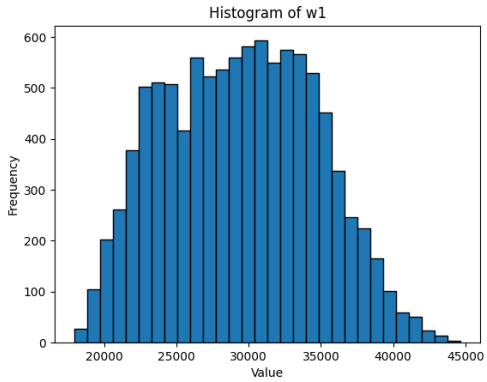

Fig. 1. Histogram of the column $w1$ of the electric demand dataset.

III. Experimental set up

A random subset of 40% of the total test instances was carefully selected to ensure that the sample accurately represents the broader distribution of the data. This approach was taken to maintain the diversity and variability inherent in the full dataset, ensuring that the selected instances reflect the key characteristics and patterns.

The analysis focused on assessing the robustness of the explanations against perturbations introduced to the input data. Specifically, perturbed instances and attributes were selected at random, ensuring a diverse range of perturbations across the dataset. For the selected attributes, perturbations were applied by either increasing or decreasing their original values by 20%, meaning $\Delta = \pm 0.2$. This fact is introducing controlled variations that reflect potential changes in real-world data.

The stability of the explanations was analyzed mathematically to ensure their reliability under such conditions. Let $\phi(x)$ represent the quality metric RexQUAL generated for a given input x.

The RExQUAL metric, designed to assess the quality of explanations provided by XAI techniques, is rooted in the global evaluation of rule-based explanations. This metric focuses on quantifying the relevance, coherence, and reliability of the sets of rules generated by the XAI method. By considering the entire rule set rather than individual rules in isolation, RExQUAL provides a comprehensive measure of the explanatory power of the technique, ensuring that the insights offered are both meaningful and consistent across the dataset.

The consistency of the explanations was evaluated by examining whether the generated rules, feature importance scores, and associated metrics (global support, global confidence) remained stable across the perturbed inputs. This approach provided a quantitative framework for assessing the robustness of the XAI technique, ensuring that the explanations it produced were reliable and interpretable, even in the presence of input variability.

IV. Results

The results, presented in Table I, illustrate the impact of perturbations on several key performance metrics, including the number of generated rules, as well as the global support, confidence, and the RExQUAL metric. These metrics provide a comprehensive view of how the introduction of perturbations influences the underlying rule-based methodology. Specifically, the number of generated rules reflects the model's ability to adapt to new variations in the data, while the global support and confidence offer insights into the reliability and strength of the generated rules across the dataset. The RExQUAL metric captures the quality and relevance of the given explanations based on the extracted rules, highlighting how perturbations might affect the interpretability and robustness of the model.

Then, Figure 2 illustrates the percentage variation of RExQUAL in relation to the proportion of perturbed instances. As expected, the more instances are perturbed, the greater the variation observed in the RExQUAL metric. However, it is worth noting that this variation does not exceed 20%, which suggests a reasonable degree of stability in the model's performance even under perturbation.

In this way, while the overall trend shows an increase in RExQUAL variation as the percentage of perturbed instances grows, significant fluctuations can be observed throughout the plot. These oscillations indicate that the impact of perturbations is not uniform and may depend on the specific characteristics of the perturbed data. For instance, the peaks around 20% and 40% perturbed instances might reflect the presence of certain influential data points or patterns that amplify the effect of perturbations on the metric. In the initial range (0%-10%), the RExQUAL variation remains relatively small, highlighting that a low percentage of perturbations has a limited impact on the metric. This behavior is consistent with the expectation that minor modifications in the input data should not drastically affect the quality of explanations provided by the model.

These findings support the robustness of the explanatory framework, as the RExQUAL metric exhibits stability across a wide range of perturbation levels. This analysis reveals that while perturbations introduce variability in the number of rules and their associated metrics, the RExQUAL values remain within reasonable bounds, reflecting the metric's robustness to moderate changes in the input data. These findings confirm that RExQUAL provides reliable assessments of explanation quality, even under varying levels of data perturbation, supporting its applicability for evaluating XAI techniques in time series prediction tasks.

It is worth noting that another source of randomness in the analysis stems from the rule generation process itself, specifically when using the Apriori algorithm. As Apriori relies on a probabilistic approach to identify frequent itemsets and generate association rules, the resulting set of rules can vary with each run, even when applied to the same dataset. This randomness can influence both the number of rules generated and their associated metrics, such as support and confidence. Despite this, the robustness of the RExQUAL metric ensures that such random variations do not significantly impact the quality of the explanations, further validating its utility for evaluating the effectiveness of XAI techniques in dynamic settings.

TABLE I
RESULTS OF STABILTY ANALYSIS.

Perturb Instances	% Perturb Instances	Perturb Attributes	% Pert Attributes	K	F	Num Rules	Global Sup	Global Conf	RExQUAL	RExQUAL Variation	% RExQUAL Variation
0	0.00	0	0	110	20	11	0.328	0.327	0.0590	0.0000	0.000
5	0.19	5	25	86	20	11	0.318	0.329	0.0574	0.0015	2.627
10	0.38	10	50	134	20	12	0.318	0.323	0.0617	0.0027	4.380
15	0.57	15	75	122	20	11	0.325	0.326	0.0582	0.0007	1.285
20	0.76	20	100	119	20	10	0.335	0.335	0.0560	0.0029	5.201
25	0.95	20	100	114	20	11	0.331	0.332	0.0604	0.0014	2.385
30	1.14	20	100	104	20	10	0.330	0.330	0.0545	0.0044	8.116
35	1.32	20	100	115	20	9	0.371	0.377	0.0629	0.0040	6.297
40	1.51	20	100	119	20	12	0.300	0.300	0.0539	0.0050	9.286
45	1.70	20	100	123	20	11	0.321	0.322	0.0570	0.0020	3.488
50	1.89	20	100	163	20	12	0.320	0.320	0.0615	0.0026	4.208
100	3.79	20	100	165	20	10	0.333	0.334	0.0556	0.0034	6.119
150	5.68	20	100	232	20	11	0.332	0.337	0.0616	0.0026	4.249
200	7.57	20	100	258	20	12	0.322	0.322	0.0623	0.0033	5.319
250	9.46	20	100	299	20	14	0.295	0.295	0.0611	0.0021	3.460
300	11.36	20	100	291	20	13	0.283	0.284	0.0522	0.0067	12.900
350	13.25	20	100	345	20	11	0.334	0.334	0.0615	0.0025	4.107
400	15.14	20	100	341	20	11	0.358	0.358	0.0705	0.0116	16.400
450	17.03	20	100	430	20	12	0.288	0.288	0.0496	0.0093	18.743
500	18.93	20	100	440	20	11	0.318	0.313	0.0548	0.0042	7.617
550	20.82	20	100	411	20	12	0.289	0.289	0.0502	0.0088	17.503
600	22.71	20	100	465	20	12	0.335	0.335	0.0675	0.0085	12.629
650	24.60	20	100	292	20	13	0.277	0.277	0.0498	0.0092	18.478
700	26.50	20	100	461	20	11	0.343	0.343	0.0647	0.0057	8.886
750	28.39	20	100	472	20	10	0.380	0.380	0.0722	0.0133	18.351
800	30.28	20	100	510	20	11	0.361	0.361	0.0717	0.0128	17.789
850	32.17	20	100	635	20	8	0.388	0.388	0.0602	0.0012	2.002
900	34.07	20	100	647	20	9	0.401	0.401	0.0723	0.0133	18.458
950	35.96	20	100	443	20	9	0.386	0.386	0.0669	0.0080	11.930
1000	37.85	20	100	384	20	10	0.367	0.367	0.0673	0.0083	12.347
1100	41.64	20	100	167	20	11	0.297	0.297	0.0486	0.0104	21.313

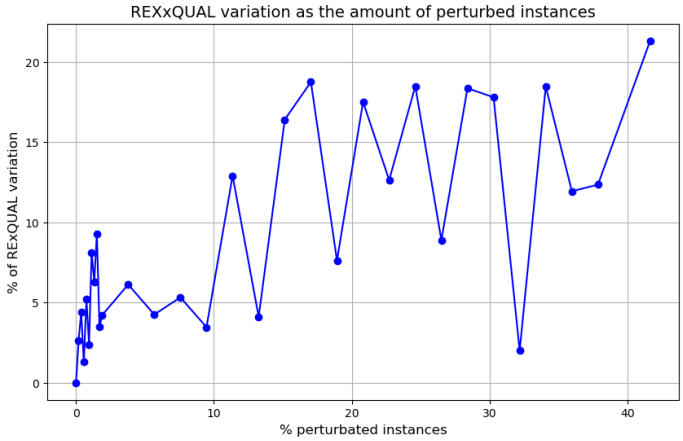

Fig. 2. Evolution of the variation of RExQUAL according to the amount of perturbated instances.

REFERENCES

[1] S. O. Hansson, and G. Helgesson, G., "What is stability?," *Synthese*, vol. 136, 2003, pp. 219–235.

[2] R. Talavera, R. Pérez-Chacón, M. Martínez-Ballesteros, A. Troncoso, and F. Martínez-Álvarez, "A nearest neighbours-based algorithm for big time series data forecasting," *Lecture Notes in Computer Science*, vol. 5391, 2016, pp. 674–679.