mathematics

Article

Assessment of Variability in Irregularly Sampled Time Series: Applications to Mental Healthcare

Pablo Bonilla-Escribano [1,*], David Ramírez [1], Alejandro Porras-Segovia [2] and Antonio Artés-Rodríguez [1]

[1] Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Leganés, Spain and Gregorio Marañón Health Research Institute, 28911 Madrid, Spain; david.ramirez@uc3m.es (D.R.); antonio@tsc.uc3m.es (A.A.-R.)

[2] Department of Psychiatry, IIS Fundación Jiménez Díaz, 28040 Madrid, Spain; alexposeg@gmail.com

* Correspondence: pbonilla@ing.uc3m.es

Abstract: Variability is defined as the propensity at which a given signal is likely to change. There are many choices for measuring variability, and it is not generally known which ones offer better properties. This paper compares different variability metrics applied to irregularly (nonuniformly) sampled time series, which have important clinical applications, particularly in mental healthcare. Using both synthetic and real patient data, we identify the most robust and interpretable variability measures out of a set 21 candidates. Some of these candidates are also proposed in this work based on the absolute slopes of the time series. An additional synthetic data experiment shows that when the complete time series is unknown, as it happens with real data, a non-negligible bias that favors normalized and/or metrics based on the raw observations of the series appears. Therefore, only the results of the synthetic experiments, which have access to the full series, should be used to draw conclusions. Accordingly, the median absolute deviation of the absolute value of the successive slopes of the data is the best way of measuring variability for this kind of time series.

Keywords: ecological momentary assessment (EMA); Hawkes process; irregularly sampled time series; variability

Citation: Bonilla-Escribano, P.; Ramírez, D.; Porras-Segovia, A.; Artés-Rodríguez, A. Assessment of Variability in Irregularly Sampled Time Series: Applications to Mental Healthcare. *Mathematics* **2021**, *9*, 71. https://dx.doi.org/10.3390/math9010071

Received: 7 December 2020
Accepted: 24 December 2020
Published: 31 December 2020

1. Introduction

Variability, also known as statistical dispersion, scatter, or spread, can be defined as the propensity at which a given signal is likely to change. The analysis of variability is of major importance in many fields, such as astrophysics [1,2], hydrology [3], agriculture [4], or ecology [5,6]. Measuring variability also has important applications in medicine. For example, it can be used to accurately measure the bioequivalence of doses of different drugs [7], predict medication demand in a hospital [8], or to measure heart rate variability, which in turn can be used as a biomarker of anxiety [9]. Variability is of particular usefulness in mental healthcare, as fluctuations in suicidal ideation have been identified as a phenotypic marker for stress-induced suicide risk [10]. In this paper, we focus on how to mathematically describe variability in nonuniformly sampled time series, which can be used in turn to alleviate problems in mental healthcare. It must be mentioned that there is a surge in the application of statistics and machine learning to solve other problems in medicine, such as classifying normal versus cancerous cells [11] or diagnosing mental disorders with electroencephalogram data [12]. For a discussion about how these methods should be incorporated in practice, refer to the work in [13].

A variability measure is an operator that transforms a time series into a single scalar, whose value hinges on the dispersion of the values of the sample. In order to make them easier to understand, many of those measures are expressed in the same units as the studied data and are non-negative, in such a way that a value of 0 indicates no change at all. Most of the measures commonly used in psychiatry have these characteristics. One possible way to achieve this is to summarize the expected distance with respect to a reference point [14]. For instance, if the reference point is the sample mean, and the expected distance is measured as the square root of the sum of the squared distances to the mean, divided by the sample size minus one, the dispersion will be expressed in terms of standard deviations. However, if the median and the median of the absolute distances are the selected criteria, the dispersion will be measured in terms of median absolute deviations, which can be more robust to outliers.

Another approach to express variability in the same units of the observations is to compute the difference of some sorted observations, thereby removing the need of a reference value. This can be useful when such a computation can get skewed by the inherent noise or other factors. In particular, if the difference is taken from the most extreme values, variability is measured in terms of range, but it is possible to compute it making use of the different percentiles of the dataset. In other cases, it may be convenient to measure variability independently of both the particular scale and the units, so that different datasets can be compared [15]. The simplest way to achieve this is to obtain ratios of variability measures, for example, the ratio of the standard deviation and the mean expresses variability as coefficients of variation [16].

The analysis of variability requires the definition of a variable of interest whose changes are computed. In psychiatry, this variable of interest is usually a symptom, which has traditionally been measured by questionnaires administered in clinical sites. This method of assessment presents several disadvantages, such as recall bias (difficulty to remember past events) [17] or ecological invalidity (by asking the questions outside the usual environment) [18]. To alleviate this, a research line has tried to obtain extemporaneous feedback from the participants by regularly requesting them to answer a set of questions via their electronic devices. This approach, which can be categorized as ecological momentary assessment (EMA) [19], suffers from fatigue effects. This consists in a decrease in the number of questions answered as time goes by, which can result in withdrawal from the study [20].

To decrease such fatigue, our research group has recently adopted a new sampling approach for the questions that consists in randomly asking some of them out of a fixed pool, thereby decreasing the sense of repetitiveness [21]. This approach results in studies with longer follow-up periods, at the expense of every set of responses for the same question being sampled non-equidistantly (nonuniformly). Because of this, special care must be taken when measuring variability when using this approach.

In this paper, we study and propose metrics which are suited for this kind of data. However, the findings of this research are not only useful for this specific type of design, as they are also applicable to any analysis in which a significant amount of missing data is produced, yielding a nonuniformly sampled time series. Missing data represent a particularly common issue in EMA studies and it is usually dealt with discarding data points that do not meet certain criteria, such as removing responses provided over 30 min after being prompted [22]. We expect that our findings will provide an accurate method of analyzing variability in non-equidistantly (nonuniformly) sampled time series, what will also allow to better cope with missing data in mental health time series.

The rest of the paper is organized as follows. Section 2 summarizes the main variability measures that have been used in the literature. As most of these measures have only been applied to equidistant, dense time series, in Section 3 we propose some modifications to those measures so that they can be applied to irregularly sampled data. In this section, we also present some novel methods of measuring variability that could also be advantageous for this task. Then, Section 4 analyzes these methods using synthetic data, while Section 5 makes use of real data to assess their robustness. To correctly interpret the results, Section 6 provides additional simulations. Finally, in Section 7 we summarize our findings, discuss their implications, and we point to next lines of research.

2. Review of Variability Assessment

In this section, we summarize the most important measures used for assessing variability in mental healthcare time series. One of the most popular measures is the *root mean square of successive differences* (RMSSD) [10], that is,

$$\text{RMSSD} = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T-1}(x_{t+1}-x_t)^2}, \tag{1}$$

where T is the number of observations, and x_t is the time series value, for instance, the response of the patient recorded at (discrete) time t in the considered application. The square is taken so that increases and decreases are not mutually canceled. This measure assumes that changes in either direction count the same. Such an assumption is usually fulfilled in psychiatry research, as commonly used scales are sometimes measured so that higher scores indicate better health status, and sometimes so that higher scores indicate worse health status [23].

The RMSSD contains some information about the temporal evolution of the assessments, as it is only invariant to complete reversal and some pairwise random permutations of the observations. This measure, normalized by the mean or by the standard deviation of the sample, also known as *coefficient of variation* (CoV) and *ratio of determination* (RoD), respectively, has been explored in [24] to produce variability measurements less sensitive to the mean value of the observations. Finally, notice that the square root is taken to recover the natural units of the explored variable in the series, though some studies do not do it [25]. This choice is not important, since the square root is a monotone transformation: it is similar to working with standard deviations or variances.

Another relevant measure of variability is the *standard deviation* (SD) of the observations [26]. This measure has the advantage of providing an interpretable parametric meaning in the case that the observations are normally distributed. However, this method does not carry any information about the order of the data (any permutation yields the same SD), and this approach is not appropriate when the distribution of the data contains many outliers or it is highly nonsymmetric [27]. A related measure that has also been employed is the *standard error of the mean* (SEM) [28], which is given by the SD divided by the square root of the number of observations.

Although simple and nonlinear, the *range of the observations*, measured as the difference between the most extreme values, has also been used [29]. On the other hand, more refined measures, such as the *entropy* (H) or the Teager–Kaiser energy operator (TKEO) have also been explored [30]. The former comes from the Information Theory field and it is a measure of data uncertainty, but it is not sensitive to local variations. We shall estimate it as

$$\text{H} = -\sum_{\substack{\{i=1,\ldots,N\} \\ \setminus\{i:\hat{p}(r_i)=0\}}} \hat{p}(r_i)\cdot\log_2\left(\hat{p}(r_i)\right), \tag{2}$$

where $\mathcal{R} = \{r_1,\ldots,r_N\}$ is the set of N possible values in the time series and $\hat{p}(r_t)$ is the estimated probability of r_t, which is given by $\hat{p}(r_t) = \frac{\#\left(r_t=\{x_t\}_{t=1}^{T}\right)}{T}$, with $\#(\cdot)$ being the count function. Notice that (2) is measured in bits, and the sum explicitly removes terms involving null probabilities, which otherwise would create numerical indeterminations. Indeed, it is judicious to do so, as one of the properties of the entropy is that incorporating or discarding events with zero probability does not alter its value. The TKEO is an approximation of the energy of a signal, which depends on its amplitude and frequency. As highly variable signals have high frequency components, larger TKEO values are associated with greater variability. This metric requires at least three data points to be computed, and it may produce negative values under some circumstances [31]. To avoid this, we will use the *mean of the absolute value of the TKEO* (MATKEO) as in [32], that is,

$$\text{MATKEO} = \frac{1}{T-2}\sum_{t=2}^{T-1}\Big|x_t^2 - (x_{t-1}\cdot x_{t+1})\Big|. \tag{3}$$

One final comment is in order. The variance terms of mixed-effect models have been used as an alternative approach to measure variability, either with raw [33] or with grand mean centered data [34]. However, we do not further analyze this model-based approach, as it is unfeasible to make fair comparisons of models tailored for specific settings, and these methods are somehow equivalent to computing the standard deviation of the data.

3. Variability Assessment of Irregularly Sampled Data

The variability measures presented above were not designed to cope with irregularly (nonuniformly) sampled time series. Therefore, in this section we propose some modifications to adapt them to such setting. First, notice that if (1) were directly applied to nonuniformly sampled data, its estimation would produce a non-desired result, as the same change in the scores, e.g., 20 units, would weight the same regardless of the time that it took such a change to happen, e.g., one day or one month. To correct this, we argue that the RMSSD in (1) can be understood as a finite-difference approximation of the square root of the mean of the squared first derivative, that is,

$$\text{RMSSD} = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T-1}(x_{t+1}-x_t)^2} \approx \sqrt{\frac{1}{t_N - t_1}\int_{t_1}^{t_N}\left(\frac{dx_t}{dt}\right)^2 dt}, \tag{4}$$

where t_1 and t_N are the times at which the first and the last observations are taken, respectively. Based on this observation, more refined approximations of the integral can be used, for instance,

$$\sqrt{\tfrac{1}{t_N-t_1}\int_{t_1}^{t_N}\left(\tfrac{dx_t}{dt}\right)^2 dt} \approx \sqrt{\tfrac{1}{t_N-t_1}\sum_{t=1}^{T-1}\left(\tfrac{x_{t+1}-x_t}{h_t}\right)^2 h_t} = \sqrt{\tfrac{1}{t_N-t_1}\sum_{t=1}^{T-1}\tfrac{(x_{t+1}-x_t)^2}{h_t}}, \tag{5}$$

which is a finite-difference approximation with (possibly) non-homogeneous step h_t between x_{t+1} and x_t. Moreover, (5) reduces to (1) when $h_t = 1,\ \forall t = 1,\ldots,T-1$, in which case $\frac{t_N-t_1}{T-1} = 1$ as well. This measure, which we call the *root mean square of successive slopes* (RMSSS), is expressed in units of change in the data per unit of time. Building on this interpretation, we can propose measures that take into account the sampling interval, i.e., based on slopes. Specifically, we propose to work with the square root of the squared slopes of successive samples, that is, the absolute value. Then, we can use the measures presented in the previous section and apply them to absolute-value slopes (therefore, we shall append the subscript "as" to the abbreviations of those measures when there is risk of confusion) instead of the raw observations (the acronyms of which will end by the subscript "raw").

We introduce other measures for summarizing the time series into a scalar, which have been useful in other fields. In particular, we shall explore the potential of (1) the *interquartile range* (IQR) [35], that is, the value of the 75th percentile minus that of the 25th one. (2) The *Gini's mean difference* (GMD) [36], namely, the expected value of the absolute difference of every possible pair of the observed values,

$$\text{GMD} = \frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T}|x_t - x_{t'}|. \tag{6}$$

Moreover, (3) the *median absolute deviation* (MAD) [37], given by

$$\text{MAD} = \text{median}\Big(\big\{|x_t - \tilde{x}|\big\}_{t=1}^{T}\Big), \tag{7}$$

where median$(\cdot)$ is the median operator and $\tilde{x} = \text{median}\left(\{x_t\}_{t=1}^{T}\right)$.

Notice that for clarity of presentation, (6) and (7) compute the corresponding measures for the raw scores, as the previous equations, but the variability measures can easily be assessed in terms of the absolute value of the slopes. For instance, the GMD of the absolute value of the slopes is given by

$$\frac{1}{(T-1)^2} \sum_{t=1}^{T-1} \sum_{t'=1}^{T-1} |z_t - z_{t'}|, \tag{8}$$

where $z_t = |\frac{x_{t+1}-x_t}{h_t}|$. For an schematic view of all the measures analyzed in this work, refer to Figure 1.

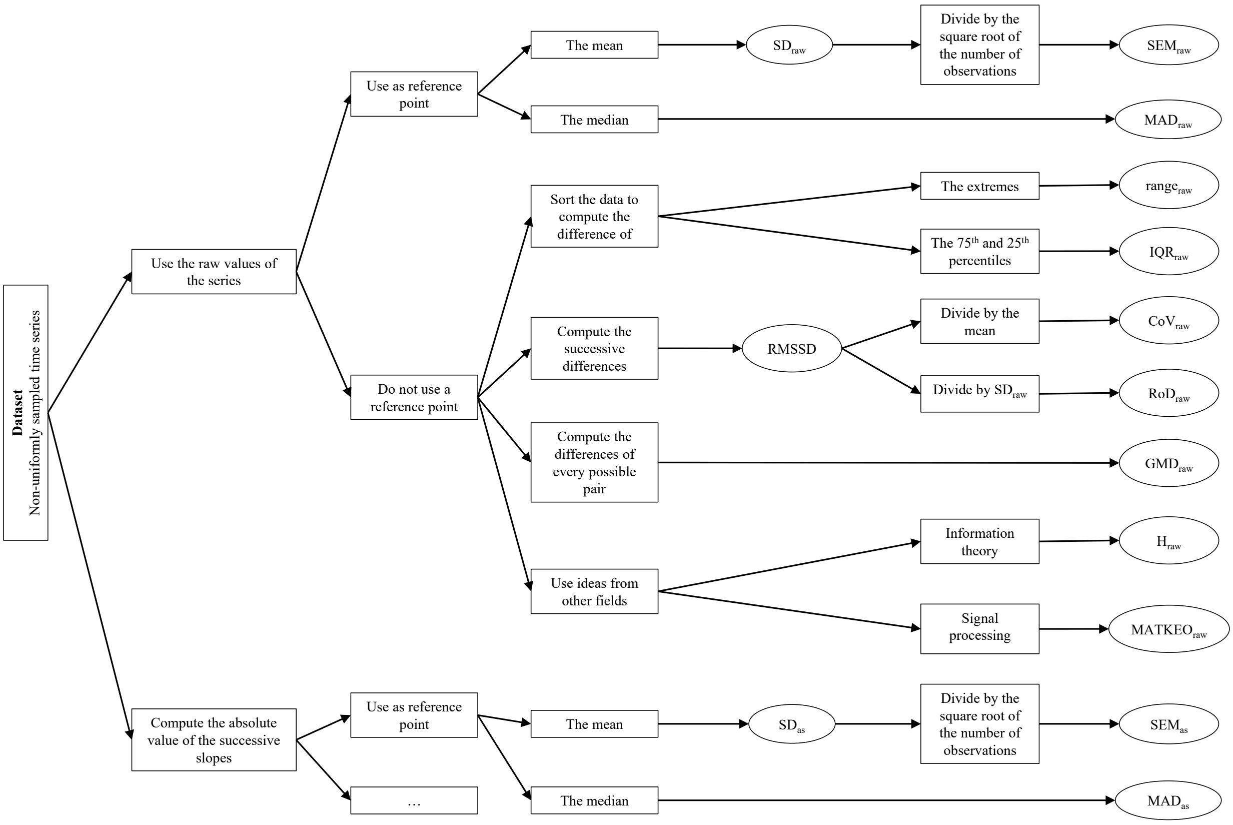

Dataset
Non-uniformly sampled time series
Use the raw values of the series
Use as reference point
The mean
SD$_{raw}$
Divide by the square root of the number of observations
SEM$_{raw}$
The median
MAD$_{raw}$
Do not use a reference point
Sort the data to compute the difference of
The extremes
range$_{raw}$
The 75th and 25th percentiles
IQR$_{raw}$
Compute the successive differences
RMSSD
Divide by the mean
CoV$_{raw}$
Divide by SD$_{raw}$
RoD$_{raw}$
Compute the differences of every possible pair
GMD$_{raw}$
Use ideas from other fields
Information theory
H$_{raw}$
Signal processing
MATKEO$_{raw}$
Compute the absolute value of the successive slopes
Use as reference point
…
The mean
SD$_{as}$
Divide by the square root of the number of observations
SEM$_{as}$
The median
MAD$_{as}$

Figure 1. Block diagram of the analyzed variability measures. The ellipsis is used to prevent repetition of the blocks for the metrics based on the absolute value of the successive slopes. The meaning of all quantities can be found in the table of abbreviations.

One final comment is in order. Overall, the complexity of the studied variability measures is comparable, since most measures are $\mathcal{O}(T)$ computationally complex, where $\mathcal{O}(\cdot)$ denotes the big O notation [38]. All the metrics based on the absolute value of the slopes requires the additional computation of the absolute value of the successive slopes, but this is an inexpensive operation, which requires a number of operations which increases linearly with the number of observations. On the other hand, the (exact) sorting of the observed values used to compute some metrics, like IQR, is $\mathcal{O}(T \log T)$ computationally complex [39]. Further, normalized metrics (i.e., SEM, CoV, and RoD) need to iterate twice the time series, but this can be done parallel. Finally, notice that the complexity of computing H$_{\text{raw}}$ also increases linearly as the number of possible values the time series can take on, N, is increased. The only metric that could be computationally more demanding is GMD, which has quadratic complexity in the number of observations, as can be seen in (6). However, for time series with few to moderate observations, the differences in time complexity are negligible in practice.

4. Comparison on Synthetic Data

In this section, we study several variability measures using synthetic data. To the best of our knowledge, there is no previous work that has simulated responses to mental health questionnaires to assess it. Therefore, we describe a possible way to achieve this, which is supported by the item response theory (IRT). The IRT states that the answers that a person provides to a questionnaire can be modeled by a transformation of a given latent variable [40], which we shall denote by Ψ. Without loss of generality, let $\Psi \in \{\psi \in \mathbb{R} : 0 \leq \psi \leq 100\}$. For instance, if Ψ measures the participants' mood, 100 may correspond with an objective, saturated feeling of elation, while 0 would be the worst possible feeling of sorrow and any value in between them represents an (objective) feeling proportional to the distance to the extremes. In our experiments, we set the possible responses to $\mathcal{R} = \{x \in \mathbb{Z} : 0 \leq x \leq 100\}$. In doing so, we account for the measurement error given by the precision of the scale. However, it is important to model other sources of noise, such as respondent inconsistency [41]. Indeed, it would be very unlikely that a participant always reports the same value in the questionnaire for the same value of Ψ. Furthermore, different participants, all sharing the same latent value of Ψ, may provide higher or lower values in the questionnaire if they tend to be optimistic or pessimistic, respectively.

Defining t as time in days, we exemplify three different changes for the latent state over 60 days as follows,

$$\Psi_t = \frac{100}{60}t \tag{9}$$

$$\Psi_t = 20\cos\left(\frac{2\pi t}{7}\right) + 50 \tag{10}$$

$$\Psi_t = 40\cos\left(\frac{2\pi t}{7}\right) + 50. \tag{11}$$

Thus, (9) represents a slow linear drift towards the maximum value, while (10) and (11) model oscillations around the medium value with a periodicity of one week. Notice that (11) has double the amplitude of (10). These are prototypic types of behaviors that were found to be good examples of other more general cases which were analyzed but are not shown here for simplicity. Those other more general cases include a cosine with increasing amplitude or abrupt ground levels changes as time goes on, and a cubic function. In all those cases we arrived to the same conclusions.

Another limitation that should be considered is the temporal resolution. In our experiments, we restrict the observations to a maximum of one per day. Consequently, the first step is to obtain the days when the data will be acquired by sampling uniformly without replacement the set of 60 possible days. Then, using Equation (9), (10), or (11), the true value of the latent state is obtained. To account for the respondent inconsistency, we model the answer the participants would be willing to provide as a random realization of a normal distribution with mean Ψ_t and standard deviation SD_n, as normal distributions are well-suited for modeling responses to Likert questionnaires [42]. Finally, to account for the error produced by the precision of the scale, the observed data will be projected onto the set $\mathcal{R}$.

In order to objectively assess the variability measures, we follow an approach similar to that of [43]. That is, we evaluate the measures in a favorable scenario, and then we obtain the performance degradation by computing the Cohen's d between the observed values in the favorable scenario and those observed in unfavorable scenarios. In particular, we compute it as

$$\text{Cohen's d} = \frac{\overline{x}_{ref} - \overline{x}_{comp}}{\sqrt{\frac{(N_{ref}-1)\times(SD_{ref})^2+(N_{comp}-1)\times(SD_{comp})^2}{N_{ref}+N_{comp}-2}}}, \tag{12}$$

where $\overline{x}_{ref}$ is the mean in the reference set, in other words, the mean of a given variability measure in the favorable scenario. Moreover, $\overline{x}_{comp}$ represents the mean of the comparison set, i.e., the mean of the same variability measure in the unfavorable scenario, and N_{ref} (conversely N_{comp}) is the number of observations in the reference (conversely, comparison) dataset. Finally, SD_{ref} (conversely SD_{comp}) stands for the standard deviation in the reference (conversely, comparison) dataset.

We shall complicate the problem by reducing the number of observations and increasing the variance of the Gaussian function for the same Ψ_t. In the favorable scenario, we observe the 60 possible answers with very low respondent inconsistency, i.e., $SD_n = 0.5$. Figure 2 shows one realization under such favorable settings for the three explored changes, and the mean values of the variability measures over 4000 realizations are presented in Figures 3a, 4a and 5a. Notice that in these figures, to facilitate the interpretation, the same colormap has been used, which spans the range of all the data in the three figures. Finally, the performance degradation is depicted in Figures 3b, 4b and 5b, also using a single colormap, which spans the full range of the data in these three new figures. To further ease the interpretation, Figures 3b, 4b and 5b show the *absolute value* of the Cohen's d, as it is of little relevance whether the measure underestimates or overestimates the variability in the complicated scenario: for the robustness analysis the importance lies in the magnitude of the discrepancy. As previously explained, each row in these figures corresponds to a different value of SD_n and number of observations, N_{comp}, where this reduced number of observations is obtained by randomly sampling the 60 days without replacement.

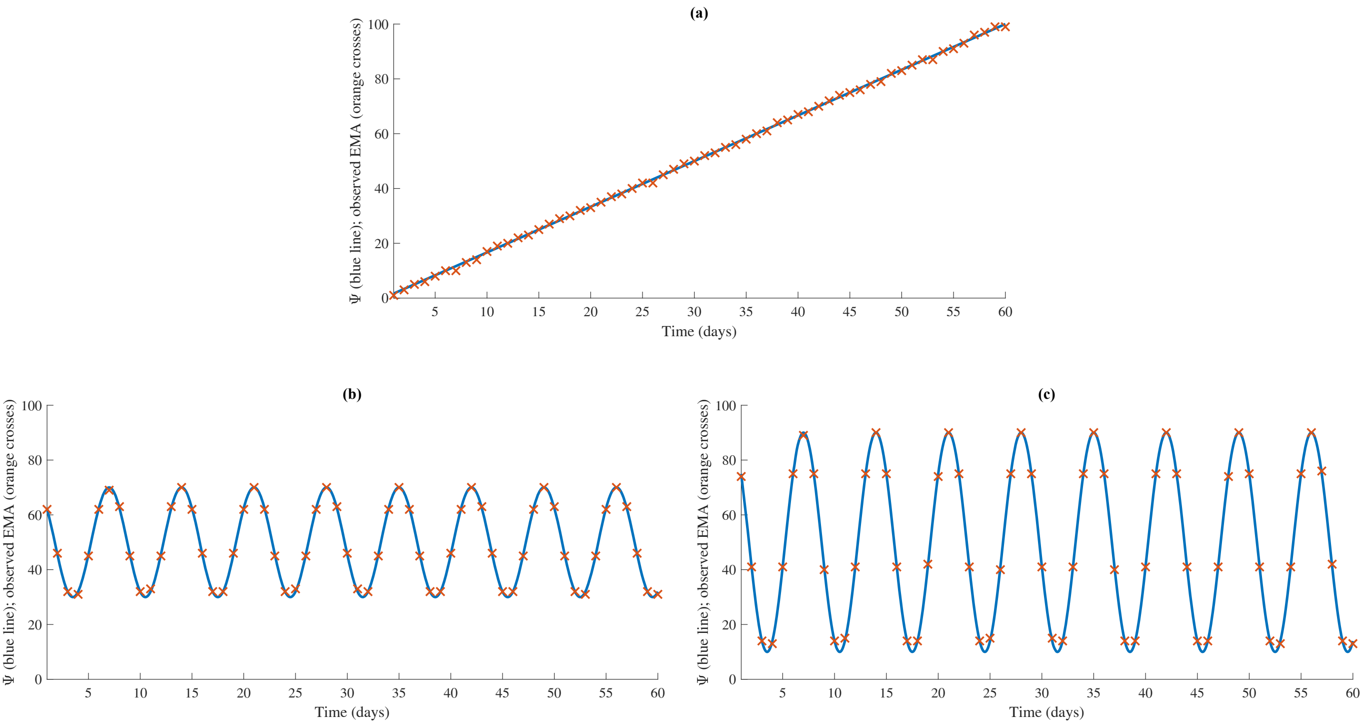
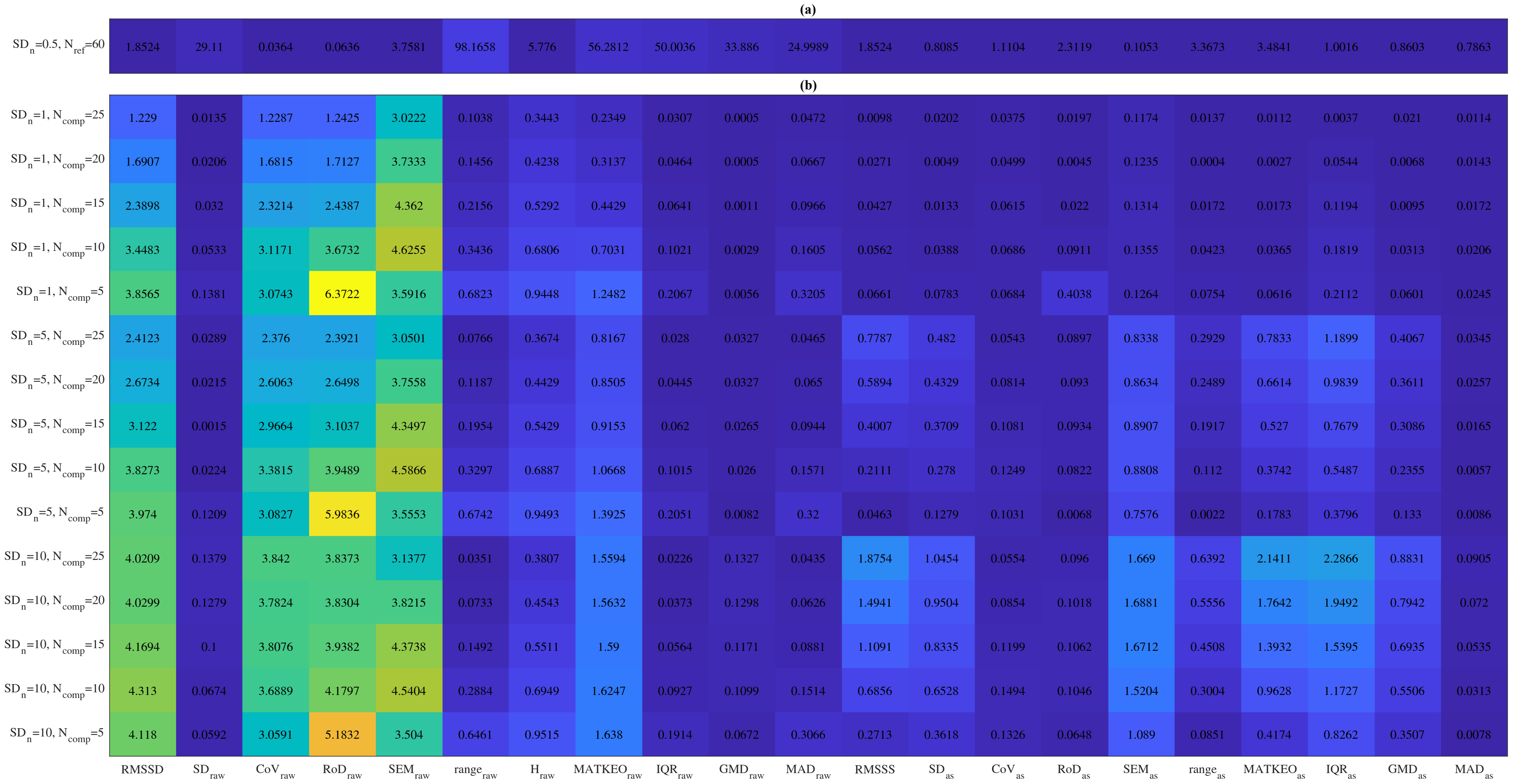


Figure 2. One random realization in the favorable scenario. (**a**) Linear function. (**b**) Cosine function with amplitude 20. (**c**) Cosine function with amplitude 40.



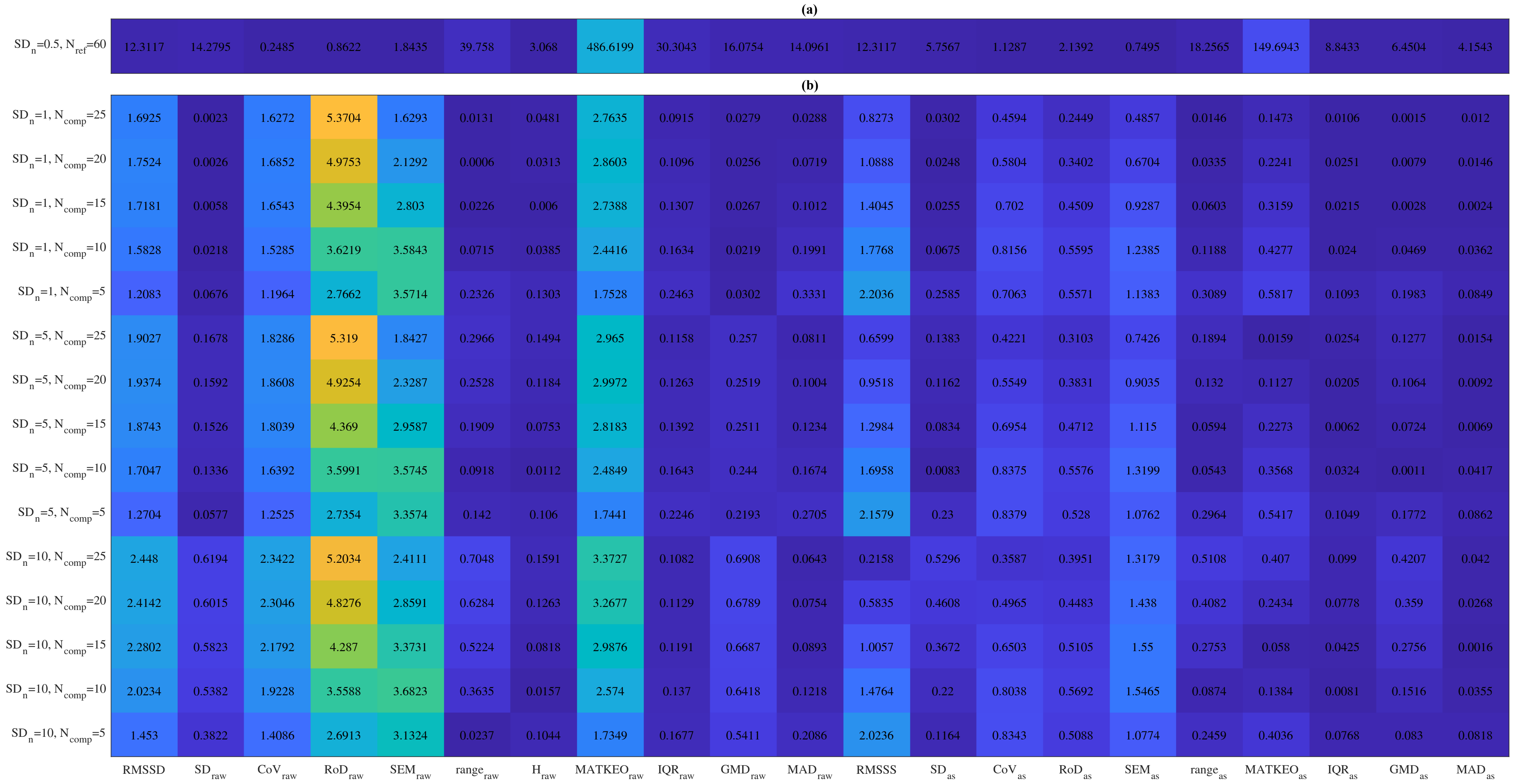
(a)

	RMSSD	SD_{raw}	CoV_{raw}	RoD_{raw}	SEM_{raw}	$range_{raw}$	H_{raw}	$MATKEO_{raw}$	IQR_{raw}	GMD_{raw}	MAD_{raw}	RMSSS	SD_{as}	CoV_{as}	RoD_{as}	SEM_{as}	$range_{as}$	$MATKEO_{as}$	IQR_{as}	GMD_{as}	MAD_{as}
$SD_{n}=0.5, N_{ref}=60$	1.8524	29.11	0.0364	0.0636	3.7581	98.1658	5.776	56.2812	50.0036	33.886	24.9989	1.8524	0.8085	1.1104	2.3119	0.1053	3.3673	3.4841	1.0016	0.8603	0.7863

(b)

	RMSSD	SD_{raw}	CoV_{raw}	RoD_{raw}	SEM_{raw}	$range_{raw}$	H_{raw}	$MATKEO_{raw}$	IQR_{raw}	GMD_{raw}	MAD_{raw}	RMSSS	SD_{as}	CoV_{as}	RoD_{as}	SEM_{as}	$range_{as}$	$MATKEO_{as}$	IQR_{as}	GMD_{as}	MAD_{as}
$SD_{n}=1, N_{comp}=25$	1.229	0.0135	1.2287	1.2425	3.0222	0.1038	0.3443	0.2349	0.0307	0.0005	0.0472	0.0098	0.0202	0.0375	0.0197	0.1174	0.0137	0.0112	0.0037	0.021	0.0114
$SD_{n}=1, N_{comp}=20$	1.6907	0.0206	1.6815	1.7127	3.7333	0.1456	0.4238	0.3137	0.0464	0.0005	0.0667	0.0271	0.0049	0.0499	0.0045	0.1235	0.0004	0.0027	0.0544	0.0068	0.0143
$SD_{n}=1, N_{comp}=15$	2.3898	0.032	2.3214	2.4387	4.362	0.2156	0.5292	0.4429	0.0641	0.0011	0.0966	0.0427	0.0133	0.0615	0.022	0.1314	0.0172	0.0173	0.1194	0.0095	0.0172
$SD_{n}=1, N_{comp}=10$	3.4483	0.0533	3.1171	3.6732	4.6255	0.3436	0.6806	0.7031	0.1021	0.0029	0.1605	0.0562	0.0388	0.0686	0.0911	0.1355	0.0423	0.0365	0.1819	0.0313	0.0206
$SD_{n}=1, N_{comp}=5$	3.8565	0.1381	3.0743	6.3722	3.5916	0.6823	0.9448	1.2482	0.2067	0.0056	0.3205	0.0661	0.0783	0.0684	0.4038	0.1264	0.0754	0.0616	0.2112	0.0601	0.0245
$SD_{n}=5, N_{comp}=25$	2.4123	0.0289	2.376	2.3921	3.0501	0.0766	0.3674	0.8167	0.028	0.0327	0.0465	0.7787	0.482	0.0543	0.0897	0.8338	0.2929	0.7833	1.1899	0.4067	0.0345
$SD_{n}=5, N_{comp}=20$	2.6734	0.0215	2.6063	2.6498	3.7558	0.1187	0.4429	0.8505	0.0445	0.0327	0.065	0.5894	0.4329	0.0814	0.093	0.8634	0.2489	0.6614	0.9839	0.3611	0.0257
$SD_{n}=5, N_{comp}=15$	3.122	0.0015	2.9664	3.1037	4.3497	0.1954	0.5429	0.9153	0.062	0.0265	0.0944	0.4007	0.3709	0.1081	0.0934	0.8907	0.1917	0.527	0.7679	0.3086	0.0165
$SD_{n}=5, N_{comp}=10$	3.8273	0.0224	3.3815	3.9489	4.5866	0.3297	0.6887	1.0668	0.1015	0.026	0.1571	0.2111	0.278	0.1249	0.0822	0.8808	0.112	0.3742	0.5487	0.2355	0.0057
$SD_{n}=5, N_{comp}=5$	3.974	0.1209	3.0827	5.9836	3.5553	0.6742	0.9493	1.3925	0.2051	0.0082	0.32	0.0463	0.1279	0.1031	0.0068	0.7576	0.0022	0.1783	0.3796	0.133	0.0086
$SD_{n}=10, N_{comp}=25$	4.0209	0.1379	3.842	3.8373	3.1377	0.0351	0.3807	1.5594	0.0226	0.1327	0.0435	1.8754	1.0454	0.0554	0.096	1.669	0.6392	2.1411	2.2866	0.8831	0.0905
$SD_{n}=10, N_{comp}=20$	4.0299	0.1279	3.7824	3.8304	3.8215	0.0733	0.4543	1.5632	0.0373	0.1298	0.0626	1.4941	0.9504	0.0854	0.1018	1.6881	0.5556	1.7642	1.9492	0.7942	0.072
$SD_{n}=10, N_{comp}=15$	4.1694	0.1	3.8076	3.9382	4.3738	0.1492	0.5511	1.59	0.0564	0.1171	0.0881	1.1091	0.8335	0.1199	0.1062	1.6712	0.4508	1.3932	1.5395	0.6935	0.0535
$SD_{n}=10, N_{comp}=10$	4.313	0.0674	3.6889	4.1797	4.5404	0.2884	0.6949	1.6247	0.0927	0.1099	0.1514	0.6856	0.6528	0.1494	0.1046	1.5204	0.3004	0.9628	1.1727	0.5506	0.0313
$SD_{n}=10, N_{comp}=5$	4.118	0.0592	3.0591	5.1832	3.504	0.6461	0.9515	1.638	0.1914	0.0672	0.3066	0.2713	0.3618	0.1326	0.0648	1.089	0.0851	0.4174	0.8262	0.3507	0.0078

Figure 3. Results for 4000 realizations with a linear function. The meaning of all quantities can be found in the table of abbreviations. (**a**) Mean values in the favorable scenario. (**b**) Performance degradation.



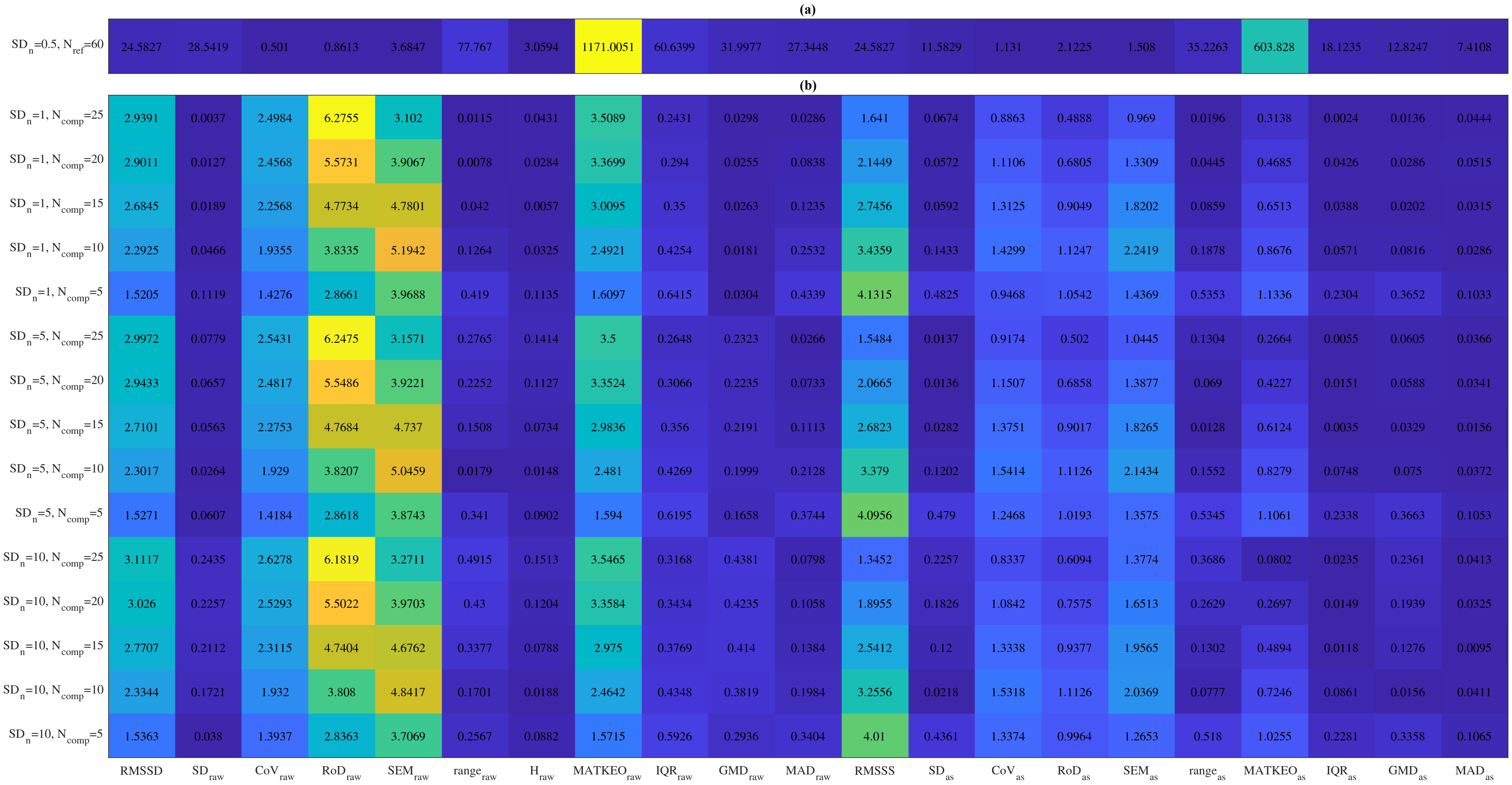
(a)

	RMSSD	SD_{raw}	CoV_{raw}	RoD_{raw}	SEM_{raw}	$range_{raw}$	H_{raw}	$MATKEO_{raw}$	IQR_{raw}	GMD_{raw}	MAD_{raw}	RMSSS	SD_{as}	CoV_{as}	RoD_{as}	SEM_{as}	$range_{as}$	$MATKEO_{as}$	IQR_{as}	GMD_{as}	MAD_{as}
SD_n=0.5, N_{ref}=60	12.3117	14.2795	0.2485	0.8622	1.8435	39.758	3.068	486.6199	30.3043	16.0754	14.0961	12.3117	5.7567	1.1287	2.1392	0.7495	18.2565	149.6943	8.8433	6.4504	4.1543

(b)

	RMSSD	SD_{raw}	CoV_{raw}	RoD_{raw}	SEM_{raw}	$range_{raw}$	H_{raw}	$MATKEO_{raw}$	IQR_{raw}	GMD_{raw}	MAD_{raw}	RMSSS	SD_{as}	CoV_{as}	RoD_{as}	SEM_{as}	$range_{as}$	$MATKEO_{as}$	IQR_{as}	GMD_{as}	MAD_{as}
SD_n=1, N_{comp}=25	1.6925	0.0023	1.6272	5.3704	1.6293	0.0131	0.0481	2.7635	0.0915	0.0279	0.0288	0.8273	0.0302	0.4594	0.2449	0.4857	0.0146	0.1473	0.0106	0.0015	0.012
SD_n=1, N_{comp}=20	1.7524	0.0026	1.6852	4.9753	2.1292	0.0006	0.0313	2.8603	0.1096	0.0256	0.0719	1.0888	0.0248	0.5804	0.3402	0.6704	0.0335	0.2241	0.0251	0.0079	0.0146
SD_n=1, N_{comp}=15	1.7181	0.0058	1.6543	4.3954	2.803	0.0226	0.006	2.7388	0.1307	0.0267	0.1012	1.4045	0.0255	0.702	0.4509	0.9287	0.0603	0.3159	0.0215	0.0028	0.0024
SD_n=1, N_{comp}=10	1.5828	0.0218	1.5285	3.6219	3.5843	0.0715	0.0385	2.4416	0.1634	0.0219	0.1991	1.7768	0.0675	0.8156	0.5595	1.2385	0.1188	0.4277	0.024	0.0469	0.0362
SD_n=1, N_{comp}=5	1.2083	0.0676	1.1964	2.7662	3.5714	0.2326	0.1303	1.7528	0.2463	0.0302	0.3331	2.2036	0.2585	0.7063	0.5571	1.1383	0.3089	0.5817	0.1093	0.1983	0.0849
SD_n=5, N_{comp}=25	1.9027	0.1678	1.8286	5.319	1.8427	0.2966	0.1494	2.965	0.1158	0.257	0.0811	0.6599	0.1383	0.4221	0.3103	0.7426	0.1894	0.0159	0.0254	0.1277	0.0154
SD_n=5, N_{comp}=20	1.9374	0.1592	1.8608	4.9254	2.3287	0.2528	0.1184	2.9972	0.1263	0.2519	0.1004	0.9518	0.1162	0.5549	0.3831	0.9035	0.132	0.1127	0.0205	0.1064	0.0092
SD_n=5, N_{comp}=15	1.8743	0.1526	1.8039	4.369	2.9587	0.1909	0.0753	2.8183	0.1392	0.2511	0.1234	1.2984	0.0834	0.6954	0.4712	1.115	0.0594	0.2273	0.0062	0.0724	0.0069
SD_n=5, N_{comp}=10	1.7047	0.1336	1.6392	3.5991	3.5745	0.0918	0.0112	2.4849	0.1643	0.244	0.1674	1.6958	0.0083	0.8375	0.5576	1.3199	0.0543	0.3568	0.0324	0.0011	0.0417
SD_n=5, N_{comp}=5	1.2704	0.0577	1.2525	2.7354	3.3574	0.142	0.106	1.7441	0.2246	0.2193	0.2705	2.1579	0.23	0.8379	0.528	1.0762	0.2964	0.5417	0.1049	0.1772	0.0862
SD_n=10, N_{comp}=25	2.448	0.6194	2.3422	5.2034	2.4111	0.7048	0.1591	3.3727	0.1082	0.6908	0.0643	0.2158	0.5296	0.3587	0.3951	1.3179	0.5108	0.407	0.099	0.4207	0.042
SD_n=10, N_{comp}=20	2.4142	0.6015	2.3046	4.8276	2.8591	0.6284	0.1263	3.2677	0.1129	0.6789	0.0754	0.5835	0.4608	0.4965	0.4483	1.438	0.4082	0.2434	0.0778	0.359	0.0268
SD_n=10, N_{comp}=15	2.2802	0.5823	2.1792	4.287	3.3731	0.5224	0.0818	2.9876	0.1191	0.6687	0.0893	1.0057	0.3672	0.6503	0.5105	1.55	0.2753	0.058	0.0425	0.2756	0.0016
SD_n=10, N_{comp}=10	2.0234	0.5382	1.9228	3.5588	3.6823	0.3635	0.0157	2.574	0.137	0.6418	0.1218	1.4764	0.22	0.8038	0.5692	1.5465	0.0874	0.1384	0.0081	0.1516	0.0355
SD_n=10, N_{comp}=5	1.453	0.3822	1.4086	2.6913	3.1324	0.0237	0.1044	1.7349	0.1677	0.5411	0.2086	2.0236	0.1164	0.8343	0.5088	1.0774	0.2459	0.4036	0.0768	0.083	0.0818

Figure 4. Results for 4000 realizations with a cosine function with amplitude 20. The meaning of all quantities can be found in the table of abbreviations. (**a**) Mean values in the favorable scenario. (**b**) Performance degradation.



(a)

	RMSSD	SD_{raw}	CoV_{raw}	RoD_{raw}	SEM_{raw}	$range_{raw}$	H_{raw}	$MATKEO_{raw}$	IQR_{raw}	GMD_{raw}	MAD_{raw}	RMSSS	SD_{as}	CoV_{as}	RoD_{as}	SEM_{as}	$range_{as}$	$MATKEO_{as}$	IQR_{as}	GMD_{as}	MAD_{as}
SD_n=0.5, N_{ref}=60	24.5827	28.5419	0.501	0.8613	3.6847	77.767	3.0594	1171.0051	60.6399	31.9977	27.3448	24.5827	11.5829	1.131	2.1225	1.508	35.2263	603.828	18.1235	12.8247	7.4108

(b)

	RMSSD	SD_{raw}	CoV_{raw}	RoD_{raw}	SEM_{raw}	$range_{raw}$	H_{raw}	$MATKEO_{raw}$	IQR_{raw}	GMD_{raw}	MAD_{raw}	RMSSS	SD_{as}	CoV_{as}	RoD_{as}	SEM_{as}	$range_{as}$	$MATKEO_{as}$	IQR_{as}	GMD_{as}	MAD_{as}
SD_n=1, N_{comp}=25	2.9391	0.0037	2.4984	6.2755	3.102	0.0115	0.0431	3.5089	0.2431	0.0298	0.0286	1.641	0.0674	0.8863	0.4888	0.969	0.0196	0.3138	0.0024	0.0136	0.0444
SD_n=1, N_{comp}=20	2.9011	0.0127	2.4568	5.5731	3.9067	0.0078	0.0284	3.3699	0.294	0.0255	0.0838	2.1449	0.0572	1.1106	0.6805	1.3309	0.0445	0.4685	0.0426	0.0286	0.0515
SD_n=1, N_{comp}=15	2.6845	0.0189	2.2568	4.7734	4.7801	0.042	0.0057	3.0095	0.35	0.0263	0.1235	2.7456	0.0592	1.3125	0.9049	1.8202	0.0859	0.6513	0.0388	0.0202	0.0315
SD_n=1, N_{comp}=10	2.2925	0.0466	1.9355	3.8335	5.1942	0.1264	0.0325	2.4921	0.4254	0.0181	0.2532	3.4359	0.1433	1.4299	1.1247	2.2419	0.1878	0.8676	0.0571	0.0816	0.0286
SD_n=1, N_{comp}=5	1.5205	0.1119	1.4276	2.8661	3.9688	0.419	0.1135	1.6097	0.6415	0.0304	0.4339	4.1315	0.4825	0.9468	1.0542	1.4369	0.5353	1.1336	0.2304	0.3652	0.1033
SD_n=5, N_{comp}=25	2.9972	0.0779	2.5431	6.2475	3.1571	0.2765	0.1414	3.5	0.2648	0.2323	0.0266	1.5484	0.0137	0.9174	0.502	1.0445	0.1304	0.2664	0.0055	0.0605	0.0366
SD_n=5, N_{comp}=20	2.9433	0.0657	2.4817	5.5486	3.9221	0.2252	0.1127	3.3524	0.3066	0.2235	0.0733	2.0665	0.0136	1.1507	0.6858	1.3877	0.069	0.4227	0.0151	0.0588	0.0341
SD_n=5, N_{comp}=15	2.7101	0.0563	2.2753	4.7684	4.737	0.1508	0.0734	2.9836	0.356	0.2191	0.1113	2.6823	0.0282	1.3751	0.9017	1.8265	0.0128	0.6124	0.0035	0.0329	0.0156
SD_n=5, N_{comp}=10	2.3017	0.0264	1.929	3.8207	5.0459	0.0179	0.0148	2.481	0.4269	0.1999	0.2128	3.379	0.1202	1.5414	1.1126	2.1434	0.1552	0.8279	0.0748	0.075	0.0372
SD_n=5, N_{comp}=5	1.5271	0.0607	1.4184	2.8618	3.8743	0.341	0.0902	1.594	0.6195	0.1658	0.3744	4.0956	0.479	1.2468	1.0193	1.3575	0.5345	1.1061	0.2338	0.3663	0.1053
SD_n=10, N_{comp}=25	3.1117	0.2435	2.6278	6.1819	3.2711	0.4915	0.1513	3.5465	0.3168	0.4381	0.0798	1.3452	0.2257	0.8337	0.6094	1.3774	0.3686	0.0802	0.0235	0.2361	0.0413
SD_n=10, N_{comp}=20	3.026	0.2257	2.5293	5.5022	3.9703	0.43	0.1204	3.3584	0.3434	0.4235	0.1058	1.8955	0.1826	1.0842	0.7575	1.6513	0.2629	0.2697	0.0149	0.1939	0.0325
SD_n=10, N_{comp}=15	2.7707	0.2112	2.3115	4.7404	4.6762	0.3377	0.0788	2.975	0.3769	0.414	0.1384	2.5412	0.12	1.3338	0.9377	1.9565	0.1302	0.4894	0.0118	0.1276	0.0095
SD_n=10, N_{comp}=10	2.3344	0.1721	1.932	3.808	4.8417	0.1701	0.0188	2.4642	0.4348	0.3819	0.1984	3.2556	0.0218	1.5318	1.1126	2.0369	0.0777	0.7246	0.0861	0.0156	0.0411
SD_n=10, N_{comp}=5	1.5363	0.038	1.3937	2.8363	3.7069	0.2567	0.0882	1.5715	0.5926	0.2936	0.3404	4.01	0.4361	1.3374	0.9964	1.2653	0.518	1.0255	0.2281	0.3358	0.1065

Figure 5. Results for 4000 realizations with a cosine function with amplitude 40. The meaning of all quantities can be found in the table of abbreviations. (**a**) Mean values in a favorable scenario. (**b**) Performance degradation.

As can be seen in Figures 3a, 4a and 5a, some measures are insensitive to the kind of change that Ψ describes. For instance, RoD_{raw} and H_{raw} provide very similar values for the two cosine functions, and COV_{as} and RoD_{as} cannot distinguish any of the three profiles. On the other hand, MAD_{raw} and GMD_{raw} are difficult to interpret, as they yield higher variability values for the slow linear drift than the periodic function, while the vast majority of the remaining measures experiment a nearly twofold increase when the amplitude of the cosine is multiplied by 2. Another issue regarding interpretability is that RoD_{raw}, RoD_{as}, and CoV_{as} are not well-defined when the data do not vary at all, because their computation would involve a division of 0 by 0 under such a situation. Similarly, CoV_raw is undetermined if all observed values are equal to zero due to the same problem. This was not observed during the simulations, as Ψ was studied under three different varying conditions and there was also noise.

With respect to the degradation in the performance as the samples become more scarce and less reliable, Figure 3b shows that measures based on the raw value of the points of the time series struggle in the simple linear drift: SEM_{raw} is the one with the highest absolute value of Cohen's d for most combinations of SD_n and N_{comp}, followed by RoD_{raw}, RMSSD and CoV_{raw} (recall that a magnitude of Cohen's d greater than 0.8 is considered a large difference [44]). In general, the rest of the variability measures based on the raw scores perform worse than those based on the absolute value of the slopes, with the notable exceptions of GMD_{raw}, which performs exceptionally well for low values of added noise, and SD_{raw} and IQR_{raw}, which offer competitive robustness. The rest of the measures based on the slopes, except for SEM_{as}, $\text{MATKEO}_{\text{as}}$, IQR_{as}, SD_{as}, and RMSSS, are robust to both added noise and missing observations. Figure 4b reveals that the worst performing measure is RoD_{raw}, followed by SEM_{raw}, when Ψ describes a cosine with amplitude 20. When we compare Figures 3b and 4b we see that even if most of the raw-based variability measures perform better with the cosine profile than in the slow drift case (especially SEM_{raw} for low values of noise and missing observations) they are still surpassed by slope-based ones, even if RMSSS, CoV_{as}, RoD_{as}, and SEM_{as} are less reliable with the periodic profile than in the linear case. When analyzing the effect of doubling the amplitude of the cosine, i.e., comparing Figures 4b and 5b, we observe that RMSSS, SEM_{raw}, and RMSSD are the measures whose performance is more affected.

When analyzing Figures 3b, 4b and 5b as a whole, we see how badly SEM_{raw}, RoD_{raw}, RMSSD, CoV_{raw}, and $\text{MATKEO}_{\text{raw}}$ systematically perform in all the cases. RMSSD is one of the less robust measures for nonuniformly sampled observations, in spite of its wide applicability with uniform sampling. On the other hand, when looking at the most reliable ways of measuring variability, we notice that CoV_{as} and RoD_{as} offer averagely robustness for the linear case, although they fail to be robust enough for the cosine changes. Therefore, they should only be used if linear changes are expected. Similarly, IQR_{as}, H_{raw}, GMD_{as}, and SD_{as} only behave reliably if the underlying change is periodic. When the sign of the the Cohen's d value is considered (recall that Figures 3b, 4b and 5b show its *absolute value*) it is observed that the measures tend to underestimate variation as the comparison set is distorted by removing observations and adding noise. When comparing the first, sixth, and eleventh rows of these figures, we can analyze the performance with the same number of observations (i.e., 25) but increasing noise levels. The raw-based variability measures are the ones more affected by noise, notably RMSSD in the linear case. In general, noise in the responses causes more distortion than missing observations, what suggests that a small, yet reliable dataset can be more informative than a larger, noisier one. Finally, only SD_{raw}, and both $\text{range}_{\text{raw}}$, MAD_{raw} and range_{as} and MAD_{as} are consistently subject to small degradations in the three cases, but MAD_{as} is the most stable one for all the cases. In light of these findings, we conclude that the MAD_{as} is the most reasonable measure to assess variability in irregularly sampled time series.

5. Comparison on Real Data

The previous section has shown the performance of the measures under three different synthetic scenarios. As the underlying data-generating process was known, it was possible to objectively characterize the robustness of each measure by computing the Cohen's d of the variability scores obtained from the *complete* time series used as a reference and a sampled version thereof. To complement that analysis, we now perform a similar experiment upon real data. However, the best reference that can be used with real data, namely, all the available responses of a participant, is already a *sample* of the complete time series. Therefore, we compute the variability measures for all the responses of all participants and use them as the reference sets. Then, up to 1000 degenerated time series are obtained by removing a given percentage of the observations in the reference sequence selecting the observations to discard uniformly without replacement at random, and the measures are computed thereon to obtain the comparison set. Finally, the Cohen's d is computed between the sets in order to assess the robustness of each variability measure.

The dataset contains the responses from the 198 participants who rated at least four times (so that the most restrictive metric, $\mathrm{MATKEO}_{\mathrm{as}}$, can be computed) each of the following items: (a) "Today I feel the wish to live" and (b) "Today I feel tired during the day because of my sleep problems." For rating the items, participants used a slider with integer precision on scale that goes from 0 to 100. The mean ±SD sequence length of the above-mentioned items among all participants are 16.79 ± 11.00 and 12.01 ± 8.01, respectively. The mean distance ±SD among subjects is 7.85 ± 9.66 and 11.17 ± 12.59 days, respectively. The results are shown in Figure 6, where the absolute value of the Cohen's d has been taken so that the results can be more easily shown on a semi-logarithmic plot. The sample is composed of patients with a history of suicidal thoughts or behaviors, and all of them gave their informed consent.

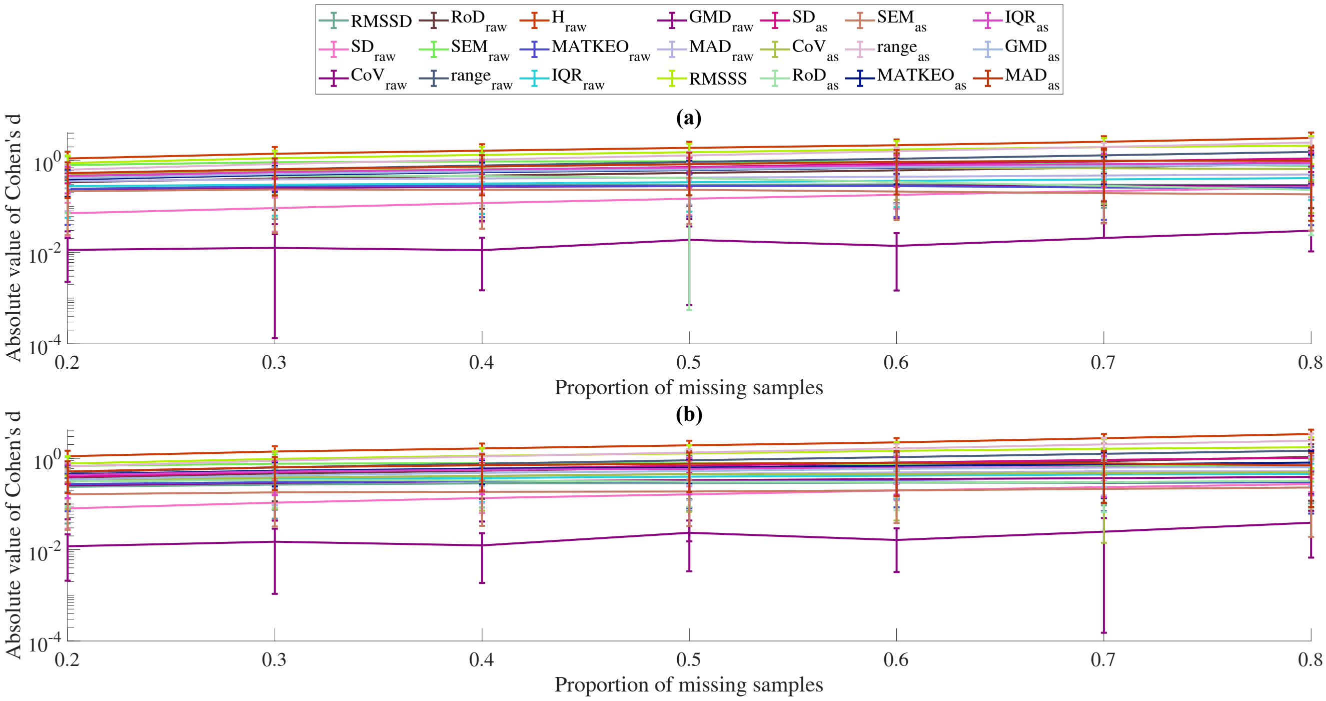


Figure 6. Performance degradation in real data for questions: (**a**) "Today I feel the wish to live." (**b**) "Today I feel tired during the day because of my sleep problems."

Overall, both Figure 6a,b exhibit a similar behavior, suggesting that the ensuing conclusions are not limited to a specific kind of questions, as the same pattern is observed for different topics (wish to live and sleep problems). Indeed, taking into account the error bars, there is not a significant loss in performance as more data are ignored. Other peculiarity of Figure 6 is that the Cohen's d values are significantly smaller than those of Figures 3b, 4b and 5b. Furthermore, according to Figure 6, GMD_{raw} is the most reliable measure, clearly outperforming the rest. Equation (6) shows that such a variability measure is very reliable to simple loss of observations (without adding noise). Indeed, the synthetic data experiments also showcased very small absolute values of Cohen's d for low levels of added noise. Finally, when analyzing in more detail Figure 6, we see that SD_{raw}, $\text{MATKEO}_{\text{raw}}$ CoV_{raw}, and RMSSD and SEM_{as} behave robustly.

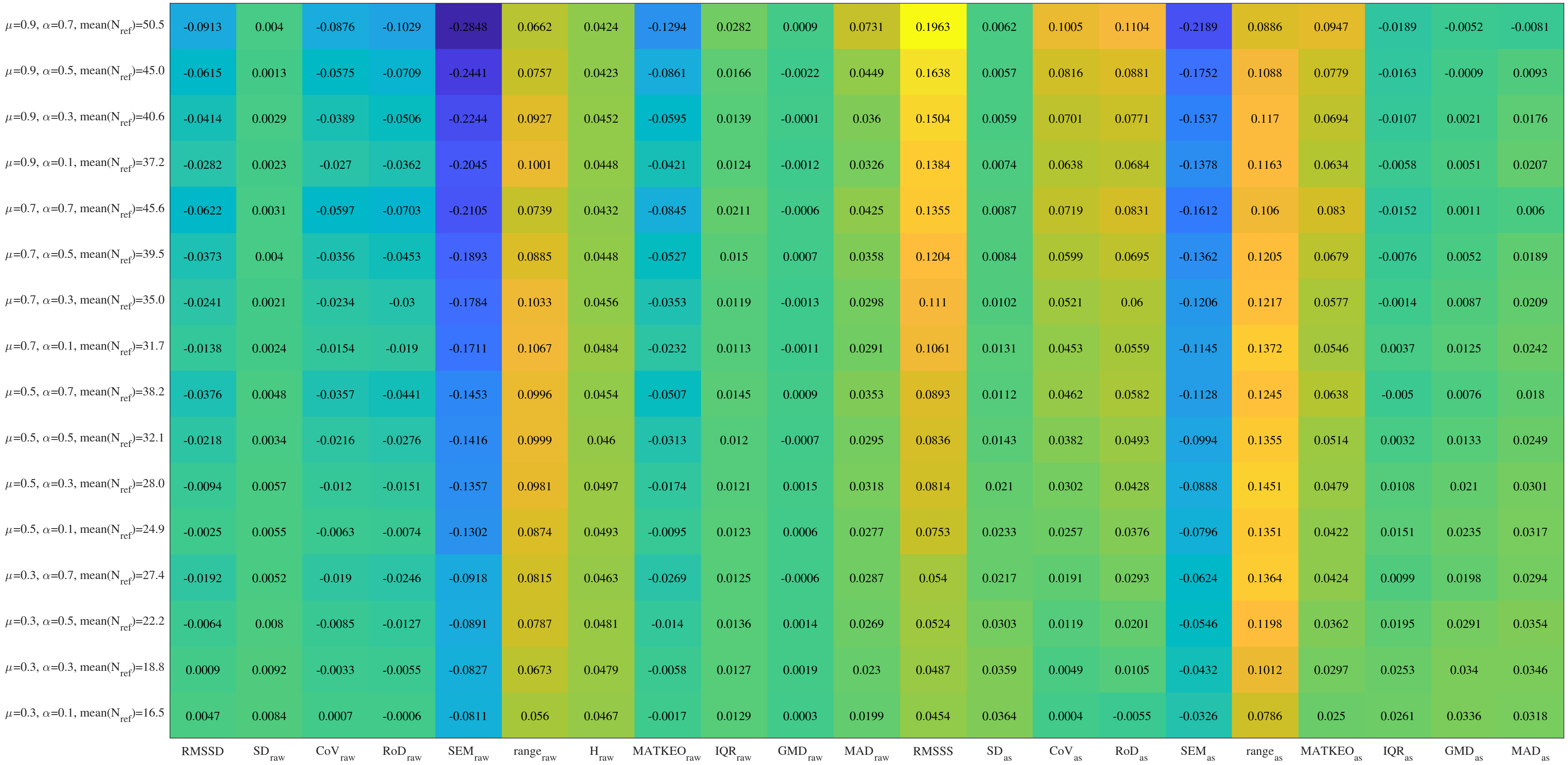
6. Additional Synthetic Data Experiment

The two previous sections have shown that there is a disagreement between the results of the synthetic data experiments and those of the real data. On one hand, synthetic experiments show that the assessment of variability becomes harder as more data are lost, and that one of the ways of obtaining robust measures is to make use of the distance among the observations by considering the absolute value of the slopes of the consecutive data points in the time series. On the other hand, results with real data depict optimistic reliability for every measure of variability, regardless of the percentage of missing observations, and they favor simple measures that are based on raw values. This discrepancy warrants an additional synthetic data analysis to ascertain if the fact that the reference used in the real data experiments is itself a *sample* of the complete time series can bias the results. For this purpose, let us consider the following experiment. First, the given set of days that will be used as a reference out of the 60 possible days is sampled according to a Hawkes process whose (conditional) intensity function is

$$\lambda^*(t) = \mu + \alpha \sum_{t_i < t} e^{-(t-t_i)}. \tag{13}$$

The intensity function measures the number of times an observation is acquired per unit of time [45]. In (13), μ is a parameter expressing the basal intensity and α controls the increase of the probability of observing additional data points after the observations have been acquired. That is, α determines the burstiness of the locations of the data points in the sequence: low values of α produce more evenly spaced time series, while high values yield points that tend to follow each other forming clusters. In this way, it is also possible to study if the relative location of the points used in the reference sequence has any influence on the results. We take this cluster-producing sampling, and the simplest way of modeling it, i.e., the Hawkes process, as a relevant example of possible kinds of structures in the location of the data points in the reference set that could induce a bias.

Second, (11) is used to simulate the time series on the sampled instances obtained using the Hawkes process in (13), yielding N_{ref} observations from which the variability measures will be evaluated, thereby obtaining the reference set. Third, 70% of those N_{ref} observations are removed by means of uniform sampling without replacement, and the variability measures are computed in order to get the comparison set. Last, the process is repeated, and the Cohen's d is obtained. Figure 7 shows those values of the Cohen's d (keeping the sign) for 4000 repetitions and different values of μ and α, which, in turn, determines the mean number of observations in the reference set, explicitly denoted by "mean(N_{ref})", for clarity. As can be seen, the magnitude of the Cohen's d in Figure 7 lies within the same range of those of the real data analysis (Figure 6), which are significantly smaller than those of the first synthetic data experiments (Figures 3b, 4b and 5b). Therefore, the low density of observations in the reference set causes a bias towards low values of Cohen's d. This stems from the fact that the lower the number of observations for the same time series, the lower are the chances of observing changes, in such a way that the comparison set becomes more and more similar to the reference one. In fact, the large majority of the variability measures attain the best results for mean sample densities as low as $(16.5 \times 100)/60 = 27.5\%$, as depicted in the bottom row of Figure 7.



	RMSSD	SD_{raw}	CoV_{raw}	RoD_{raw}	SEM_{raw}	$range_{raw}$	H_{raw}	$MATKEO_{raw}$	IQR_{raw}	GMD_{raw}	MAD_{raw}	RMSSS	SD_{as}	CoV_{as}	RoD_{as}	SEM_{as}	$range_{as}$	$MATKEO_{as}$	IQR_{as}	GMD_{as}	MAD_{as}
μ=0.9, α=0.7, mean(N_{ref})=50.5	-0.0913	0.004	-0.0876	-0.1029	-0.2848	0.0662	0.0424	-0.1294	0.0282	0.0009	0.0731	0.1963	0.0062	0.1005	0.1104	-0.2189	0.0886	0.0947	-0.0189	-0.0052	-0.0081
μ=0.9, α=0.5, mean(N_{ref})=45.0	-0.0615	0.0013	-0.0575	-0.0709	-0.2441	0.0757	0.0423	-0.0861	0.0166	-0.0022	0.0449	0.1638	0.0057	0.0816	0.0881	-0.1752	0.1088	0.0779	-0.0163	-0.0009	0.0093
μ=0.9, α=0.3, mean(N_{ref})=40.6	-0.0414	0.0029	-0.0389	-0.0506	-0.2244	0.0927	0.0452	-0.0595	0.0139	-0.0001	0.036	0.1504	0.0059	0.0701	0.0771	-0.1537	0.117	0.0694	-0.0107	0.0021	0.0176
μ=0.9, α=0.1, mean(N_{ref})=37.2	-0.0282	0.0023	-0.027	-0.0362	-0.2045	0.1001	0.0448	-0.0421	0.0124	-0.0012	0.0326	0.1384	0.0074	0.0638	0.0684	-0.1378	0.1163	0.0634	-0.0058	0.0051	0.0207
μ=0.7, α=0.7, mean(N_{ref})=45.6	-0.0622	0.0031	-0.0597	-0.0703	-0.2105	0.0739	0.0432	-0.0845	0.0211	-0.0006	0.0425	0.1355	0.0087	0.0719	0.0831	-0.1612	0.106	0.083	-0.0152	0.0011	0.006
μ=0.7, α=0.5, mean(N_{ref})=39.5	-0.0373	0.004	-0.0356	-0.0453	-0.1893	0.0885	0.0448	-0.0527	0.015	0.0007	0.0358	0.1204	0.0084	0.0599	0.0695	-0.1362	0.1205	0.0679	-0.0076	0.0052	0.0189
μ=0.7, α=0.3, mean(N_{ref})=35.0	-0.0241	0.0021	-0.0234	-0.03	-0.1784	0.1033	0.0456	-0.0353	0.0119	-0.0013	0.0298	0.111	0.0102	0.0521	0.06	-0.1206	0.1217	0.0577	-0.0014	0.0087	0.0209
μ=0.7, α=0.1, mean(N_{ref})=31.7	-0.0138	0.0024	-0.0154	-0.019	-0.1711	0.1067	0.0484	-0.0232	0.0113	-0.0011	0.0291	0.1061	0.0131	0.0453	0.0559	-0.1145	0.1372	0.0546	0.0037	0.0125	0.0242
μ=0.5, α=0.7, mean(N_{ref})=38.2	-0.0376	0.0048	-0.0357	-0.0441	-0.1453	0.0996	0.0454	-0.0507	0.0145	0.0009	0.0353	0.0893	0.0112	0.0462	0.0582	-0.1128	0.1245	0.0638	-0.005	0.0076	0.018
μ=0.5, α=0.5, mean(N_{ref})=32.1	-0.0218	0.0034	-0.0216	-0.0276	-0.1416	0.0999	0.046	-0.0313	0.012	-0.0007	0.0295	0.0836	0.0143	0.0382	0.0493	-0.0994	0.1355	0.0514	0.0032	0.0133	0.0249
μ=0.5, α=0.3, mean(N_{ref})=28.0	-0.0094	0.0057	-0.012	-0.0151	-0.1357	0.0981	0.0497	-0.0174	0.0121	0.0015	0.0318	0.0814	0.021	0.0302	0.0428	-0.0888	0.1451	0.0479	0.0108	0.021	0.0301
μ=0.5, α=0.1, mean(N_{ref})=24.9	-0.0025	0.0055	-0.0063	-0.0074	-0.1302	0.0874	0.0493	-0.0095	0.0123	0.0006	0.0277	0.0753	0.0233	0.0257	0.0376	-0.0796	0.1351	0.0422	0.0151	0.0235	0.0317
μ=0.3, α=0.7, mean(N_{ref})=27.4	-0.0192	0.0052	-0.019	-0.0246	-0.0918	0.0815	0.0463	-0.0269	0.0125	-0.0006	0.0287	0.054	0.0217	0.0191	0.0293	-0.0624	0.1364	0.0424	0.0099	0.0198	0.0294
μ=0.3, α=0.5, mean(N_{ref})=22.2	-0.0064	0.008	-0.0085	-0.0127	-0.0891	0.0787	0.0481	-0.014	0.0136	0.0014	0.0269	0.0524	0.0303	0.0119	0.0201	-0.0546	0.1198	0.0362	0.0195	0.0291	0.0354
μ=0.3, α=0.3, mean(N_{ref})=18.8	0.0009	0.0092	-0.0033	-0.0055	-0.0827	0.0673	0.0479	-0.0058	0.0127	0.0019	0.023	0.0487	0.0359	0.0049	0.0105	-0.0432	0.1012	0.0297	0.0253	0.034	0.0346
μ=0.3, α=0.1, mean(N_{ref})=16.5	0.0047	0.0084	0.0007	-0.0006	-0.0811	0.056	0.0467	-0.0017	0.0129	0.0003	0.0199	0.0454	0.0364	0.0004	-0.0055	-0.0326	0.0786	0.025	0.0261	0.0336	0.0318

Figure 7. Results for 4000 reference subsequences obtained sampling observations with a Hawkes process of a cosine of amplitude 40, and then removing 70% of the observations. The colormap spans the full range. The meaning of all quantities can be found in the table of abbreviations.

However, such a bias does not equally affect all the variability measures. Indeed, the value of Cohen's d of the measures based on the raw answers decreases faster to 0 than those based on the absolute value of the slopes, as the reference time series has fewer observations, particularly for simpler methods such as RMSSD, $\mathrm{CoV_{raw}}$, $\mathrm{RoD_{raw}}$, and $\mathrm{SEM_{raw}}$. Nonetheless, $\mathrm{MAD_{as}}$, $\mathrm{GMA_{as}}$, $\mathrm{IQR_{as}}$, and $\mathrm{SE_{as}}$, all of which performed averagely to poorly in the real data experiments, have a slight bias that increases the value of the Cohen's d as the reference losses observations. This can be explained by the fact that the absolute value of the slopes lies on a larger range of values than that of the raw observations, so the measures based on the absolute slopes can still capture some variability even when the reference observations are scant. Therefore, when the observations are removed in the comparison set, part of such variability is lost, and this discrepancy is shown in the Cohen's d between the two sets. On the other hand, normalization contributes to increase the bias towards low Cohen's d values. This explains why $\mathrm{SEM_{as}}$, $\mathrm{RoD_{as}}$, and $\mathrm{Co_{as}}$, which are based on the absolute slopes, suffer from it. In addition, $\mathrm{SEM_{raw}}$, $\mathrm{RoD_{raw}}$, and $\mathrm{Co_{raw}}$ are the measures that decrease faster to zero as the reference becomes lighter. Therefore, the division by an already biased normalizing factor, such as the SD, amplifies the bias in variability measures.

With regard to the possible influence of the relative location of the observations used as a reference, Figure 7 allows us to compare different combinations of μ and α, which have roughly the same number of mean observations in the reference set. In particular, it is possible to compare: (1) the 2nd and 5th rows to observe an increase from $\alpha = 0.5$ to $\alpha = 0.7$ with $\mathrm{mean}(N_{ref}) \approx 45$; (2) the 8th and 10th rows to see what happens when the burstiness is increased from $\alpha = 0.1$ to $\alpha = 0.5$ with $\mathrm{mean}(N_{ref}) \approx 32$; and (3) the 11th and 13th rows to study the case wherein $\alpha = 0.3$ and $\alpha = 0.7$ with $\mathrm{mean}(N_{ref}) \approx 28$. When the differences among those rows is inspected, the values are very low and no consistent pattern is found, suggesting the density of the observations used as a benchmark is what determines the bias induced when the reference time series is incomplete.

In summary, this additional synthetic data experiment shows that using incomplete time series as references induces non-negligible bias, reducing the differences between the reference and comparison sets for simple variability measures based on raw observations. Therefore, the real data results from Section 5, which have very low density of observations in the benchmark (<15%), should not be considered for analyzing the robustness of the variability measures. Instead, the synthetic data experiments from Section 4, which have access to the complete sequence, are the ones that should be taken into account.

7. Conclusions and Future Work

In this paper, we have summarized the main variability measures used in the literature, with special emphasis in mental healthcare time series. Datasets in this kind of studies have a strong tendency to be nonuniformly sampled regardless of the experimental design; even studies that aim for data uniformity end up being nonuniform due to the large amount of missing data. We have explored the ability of different commonly used variability measures to address this issue, and we have proposed and studied the potential of other variability indicators not explored yet to cope with this inimical situation. We compared a total 21 candidates, exploring their robustness and interpretability using synthetic data. Thanks to the known profiles used in the synthetic data, some variability measures, such as $\mathrm{MAD_{raw}}$ and $\mathrm{GMD_{raw}}$, have been identified as difficult to interpret, as they provided higher values for profiles that intuitively are less variable. On the other hand, these experiments identified measures based on the absolute value of the slopes of consecutive observations as the most robust ones. In particular, $\mathrm{MAD_{as}}$ was the best candidate.

However, we reached different conclusions in the experiments performed upon real data, as they showed that the most robust measures are the ones that are normalized and/or based on the raw values, such as RMSSD. To shed some light on this, an additional synthetic data experiment showed that low number of observations used in a benchmark (such as the one of the real data analysis) induces a non-negligible bias that favors the variability measures that are normalized and/or based on the raw value of the series. Therefore, only the results of the real data experiments should be used to determine which is the most reliable measure, as the full time series is used as a reference. We can therefore conclude that the best way to assess variability is computing the $\mathrm{MAD}_{\mathrm{as}}$, namely, the median absolute deviation of the absolute value of the successive slopes of the data in the time series. Finally, it would be instructive to conduct additional analyses on real and dense time series, as data density would decrease the risk of bias.

Author Contributions: Conceptualization, P.B.-E., D.R., and A.A.-R.; methodology, P.B.-E.; software, P.B.-E.; validation, D.R. and A.A.-R.; formal analysis, P.B.-E., D.R. and A.A.-R.; investigation, P.B.-E.; resources, A.A.-R.; data curation, P.B.-E.; writing—original draft preparation, P.B.-E.; writing—review and editing, P.B.-E., D.R., and A.P.-S.; visualization, P.B.-E.; supervision, D.R. and A.A.-R.; project administration, A.A.-R.; funding acquisition, A.A.-R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Ministerio de Ciencia, Innovación y Universidades under grant TEC2017-92552-EXP (aMBITION), by the Ministerio de Ciencia, Innovación y Universidades, jointly with the European Commission (ERDF), under grants TEC2017-86921-C2-2-R (CAIMAN) and RTI2018-099655-BI00 (CLARA), by the Comunidad de Madrid under grant Y2018/TCS-4705 (PRACTICO-CM), and by Fundación BBVA under project Deep-DARWIN.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of University Hospital Fundación Jiménez Díaz on the 25 June 2017 under IRB approval LSRG-1-005-16.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data are not publicly available due to ethical and privacy restrictions.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

$(\cdot)_{\mathrm{as}}$	(subscript indicating) computed upon the absolute value of the slopes
$(\cdot)_{\mathrm{n}}$	(subscript indicating) of the noise
$(\cdot)_{\mathrm{raw}}$	(subscript indicating) computed upon the (raw) scores
CoV	Coefficient of variation
EMA	Ecological momentary assessment
GMD	Gini's mean difference
H	Entropy
IQR	Interquartile range
IRT	Item response theory
MAD	Median absolute deviation
MATKEO	Mean of the absolute value of the TKEO
RMSSD	Root mean square of successive differences
RMSSS	Root mean square of successive slopes
RoD	Ratio of determination
SD	Standard deviation
SEM	Standard error of the mean
TKEO	Teager–Kaiser energy operator

References

1. Takata, T.; Mukuta, Y.; Mizumoto, Y. Modeling the variability of active galactic nuclei by an infinite mixture of Ornstein–Uhlenbeck (OU) processes. *Astrophys. J.* **2018**, *869*, 178. [CrossRef]
2. Heil, L.; Uttley, P.; Klein-Wolt, M. Power colours: Simple X-ray binary variability comparison. *Mon. Not. R. Astron. Soc.* **2015**, *448*, 3339–3347. [CrossRef]
3. Bürger, G.; Chen, Y. Regression-based downscaling of spatial variability for hydrologic applications. *J. Hydrol.* **2005**, *311*, 299–317. [CrossRef]

4. Armand, R.; Bockstaller, C.; Auzet, A.V.; Van Dijk, P. Runoff generation related to intra-field soil surface characteristics variability: Application to conservation tillage context. *Soil Tillage Res.* **2009**, *102*, 27–37. [CrossRef]
5. Köhnke, M.C.; Malchow, H. Impact of parameter variability and environmental noise on the Klausmeier model of vegetation pattern formation. *Mathematics* **2017**, *5*, 69. [CrossRef]
6. Morris, A.; Börger, L.; Crooks, E. Individual variability in dispersal and invasion speed. *Mathematics* **2019**, *7*, 795. [CrossRef]
7. Jiang, W.; Makhlouf, F.; Schuirmann, D.J.; Zhang, X.; Zheng, N.; Conner, D.; Lawrence, X.Y.; Lionberger, R. A bioequivalence approach for generic narrow therapeutic index drugs: Evaluation of the reference-scaled approach and variability comparison criterion. *AAPS J.* **2015**, *17*, 891–901. [CrossRef]
8. Silva-Aravena, F.; Ceballos-Fuentealba, I.; Álvarez-Miranda, E. Inventory management at a Chilean hospital pharmacy: Case study of a dynamic decision-aid tool. *Mathematics* **2020**, *8*, 1962. [CrossRef]
9. Kim, H.G.; Cheon, E.J.; Bai, D.S.; Lee, Y.H.; Koo, B.H. Stress and heart rate variability: A meta-analysis and review of the literature. *Psychiatry Investig.* **2018**, *15*, 235. [CrossRef]
10. Oquendo, M.A.; Galfalvy, H.C.; Choo, T.H.; Kandlur, R.; Burke, A.K.; Sublette, M.E.; Miller, J.M.; Mann, J.J.; Stanley, B.H. Highly variable suicidal ideation: A phenotypic marker for stress induced suicide risk. *Mol. Psychiatry* **2020**, 1–8. [CrossRef]
11. Hajiramezanali, E.; Imani, M.; Braga-Neto, U.; Qian, X.; Dougherty, E.R. Scalable optimal Bayesian classification of single-cell trajectories under regulatory model uncertainty. *BMC Genom.* **2019**, *20*, 1–11. [CrossRef]
12. Gore, E.; Rathi, S. Surveying machine learning algorithms on EEG signals data for mental health assessment. In Proceedings of the 2019 IEEE Pune Section International Conference (PuneCon), Pune, India, 18–20 December 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–6.
13. Wiens, J.; Saria, S.; Sendak, M.; Ghassemi, M.; Liu, V.X.; Doshi-Velez, F.; Jung, K.; Heller, K.; Kale, D.; Saeed, M.; et al. Do no harm: A roadmap for responsible machine learning for health care. *Nat. Med.* **2019**, *25*, 1337–1340. [CrossRef]
14. Siegel, A.F. Variability: Dealing with diversity. In *Practical Business Statistics*; Academic Press: Cambridge, MA, USA, 2012; pp. 95–121. Chapter 5.
15. Peña, D.; Linde, A. Dimensionless measures of variability and dependence for multivariate continuous distributions. *Commun. Stat. Theory Methods* **2007**, *36*, 1845–1854. [CrossRef]
16. Teoh, W.L.; Khoo, M.B.; Castagliola, P.; Yeong, W.C.; Teh, S.Y. Run-sum control charts for monitoring the coefficient of variation. *Eur. J. Oper. Res.* **2017**, *257*, 144–158. [CrossRef]
17. Boschloo, L.; Nolen, W.A.; Spijker, A.T.; Hoencamp, E.; Kupka, R.; Penninx, B.W.; Schoevers, R.A. The Mood Disorder Questionnaire (MDQ) for detecting (hypo) manic episodes: Its validity and impact of recall bias. *J. Affect. Disord.* **2013**, *151*, 203–208. [CrossRef]
18. Davidson, C.L.; Anestis, M.D.; Gutierrez, P.M. Ecological momentary assessment is a neglected methodology in suicidology. *Arch. Suicide Res.* **2017**, *21*, 1–11. [CrossRef]
19. Kim, J.; Nakamura, T.; Kikuchi, H.; Sasaki, T.; Yamamoto, Y. Co-variation of depressive mood and locomotor dynamics evaluated by ecological momentary assessment in healthy humans. *PLoS ONE* **2013**, *8*, e74979. [CrossRef]
20. Moitra, E.; Gaudiano, B.A.; Davis, C.H.; Ben-Zeev, D. Feasibility and acceptability of post-hospitalization ecological momentary assessment in patients with psychotic-spectrum disorders. *Compr. Psychiatry* **2017**, *74*, 204–213. [CrossRef]
21. Porras-Segovia, A.; Molina-Madueño, R.M.; Berrouiguet, S.; López-Castroman, J.; Barrigón, M.L.; Pérez-Rodríguez, M.S.; Marco, J.H.; Díaz-Oliván, I.; de León, S.; Courtet, P.; et al. Smartphone-based ecological momentary assessment (EMA) in psychiatric patients and student controls: A real-world feasibility study. *J. Affect. Disord.* **2020**, *274*, 733–741. [CrossRef]
22. Timmer, B.H.; Hickson, L.; Launer, S. Ecological momentary assessment: Feasibility, construct validity, and future applications. *Am. J. Audiol.* **2017**, *26*, 436–442. [CrossRef]
23. Hartley, J. Some thoughts on Likert-type scales. *Int. J. Clin. Health Psychol.* **2014**, *14*, 83–86. [CrossRef]
24. Choo, T.H.; Oquendo, M.A.; Stanley, B.; Galfalvy, H. RMSSD-based variability measures for suicidal ideation from EMA data. *Biol. Psychiatry* **2020**, *87*, S214. [CrossRef]
25. Babinski, D.E.; Welkie, J. Feasibility of ecological momentary assessment of negative emotion in girls with ADHD: A pilot study. *Psychol. Rep.* **2020**, *123*, 1027–1043. [CrossRef]
26. Shiyko, M.P.; Siembor, B.; Greene, P.B.; Smyth, J.; Burkhalter, J.E. Intra-individual study of mindfulness: Ecological momentary perspective in post-surgical lung cancer patients. *J. Behav. Med.* **2019**, *42*, 102–110. [CrossRef]
27. Schindler, T.M. Variability, range, interquartile range, and standard deviation. *Am. Med. Writ. Assoc. AMWA J.* **2015**, *30*, 132.
28. Bowen, R.; Baetz, M.; Hawkes, J.; Bowen, A. Mood variability in anxiety disorders. *J. Affect. Disord.* **2006**, *91*, 165–170. [CrossRef]
29. Black, A.C.; Cooney, N.L.; Justice, A.C.; Fiellin, L.E.; Pietrzak, R.H.; Lazar, C.M.; Rosen, M.I. Momentary assessment of PTSD symptoms and sexual risk behavior in male OEF/OIF/OND Veterans. *J. Affect. Disord.* **2016**, *190*, 424–428. [CrossRef]
30. Tsanas, A.; Saunders, K.; Bilderbeck, A.; Palmius, N.; Osipov, M.; Clifford, G.; Goodwin, G.; De Vos, M. Daily longitudinal self-monitoring of mood variability in bipolar disorder and borderline personality disorder. *J. Affect. Disord.* **2016**, *205*, 225–233. [CrossRef]
31. Bovik, A.C.; Maragos, P. Conditions for positivity of an energy operator. *IEEE Trans. Signal Process.* **1994**, *42*, 469–471. [CrossRef]
32. Jabloun, F.; Cetin, A.E.; Erzin, E. Teager energy based feature parameters for speech recognition in car noise. *IEEE Signal Process. Lett.* **1999**, *6*, 259–261. [CrossRef]

33. Hedeker, D.; Mermelstein, R.J.; Berbaum, M.L.; Campbell, R.T. Modeling mood variation associated with smoking: An application of a heterogeneous mixed-effects model for analysis of ecological momentary assessment (EMA) data. *Addiction* **2009**, *104*, 297–307. [CrossRef]
34. Cox, R.C.; Sterba, S.K.; Cole, D.A.; Upender, R.P.; Olatunji, B.O. Time of day effects on the relationship between daily sleep and anxiety: An ecological momentary assessment approach. *Behav. Res. Ther.* **2018**, *111*, 44–51. [CrossRef]
35. Forejt, M.; Brázdová, Z.D.; Novák, J.; Zlámal, F.; Forbelská, M.; Bienert, P.; Mořkovská, P.; Zavřelová, M.; Pohořalá, A.; Jurášková, M.; et al. Higher energy intake variability as predisposition to obesity: Novel approach using interquartile range. *Cent. Eur. J. Public Health* **2017**, *25*, 321–325. [CrossRef]
36. Yitzhaki, S. Gini's mean difference: A superior measure of variability for non-normal distributions. *Metron* **2003**, *61*, 285–316.
37. Leys, C.; Ley, C.; Klein, O.; Bernard, P.; Licata, L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *J. Exp. Soc. Psychol.* **2013**, *49*, 764–766. [CrossRef]
38. Knuth, D.E. *The Art of Computer Programming: Fundamental Algorithms*, 3rd ed.; Addison Wesley Longman Publishing Co., Inc.: Boston, MA, USA, 1997; Volume 1.
39. Chen, Z.; Zhang, A. A survey of approximate quantile computation on large-scale data. *IEEE Access* **2020**, *8*, 34585–34597. [CrossRef]
40. Allik, J. A mixed-binomial model for Likert-type personality measures. *Front. Psychol.* **2014**, *5*, 371. [CrossRef]
41. Biderman, M.D.; Reddock, C.M. The relationship of scale reliability and validity to respondent inconsistency. *Personal. Individ. Differ.* **2012**, *52*, 647–651. [CrossRef]
42. DeWees, T.A.; Mazza, G.L.; Golafshar, M.A.; Dueck, A.C. Investigation into the effects of using normal distribution theory methodology for Likert scale patient-reported outcome data from varying underlying distributions including floor/ceiling effects. *Value Health* **2020**, *23*, 625–631. [CrossRef]
43. Munoz, M.L.; van Roon, A.; Riese, H.; Thio, C.; Oostenbroek, E.; Westrik, I.; de Geus, E.J.C.; Gansevoort, R.; Lefrandt, J.; Nolte, I.M.; et al. Validity of (ultra-) short recordings for heart rate variability measurements. *PLoS ONE* **2015**, *10*, e0138921. [CrossRef]
44. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*; Academic Press: Cambridge, MA, USA, 2013.
45. Daley, D.J.; Vere-Jones, D. *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*; Springer: New York, NY, USA, 2003.