sensors

MDPI

Article

Wearables and the Quantified Self: Systematic Benchmarking of Physiological Sensors

Günther Sagl [1,*], Bernd Resch [1,2,*], Andreas Petutschnig [1], Kalliopi Kyriakou [1], Michael Liedlgruber [3] and Frank H. Wilhelm [3]

1 Department of Geoinformatics—Z_GIS, University of Salzburg, 5020 Salzburg, Austria; Andreas.Petutschnig@sbg.ac.at (A.P.); Kalliopi.Kyriakou@sbg.ac.at (K.K.)
2 Center for Geographic Analysis, Harvard University, Cambridge, MA 02138, USA
3 Department of Psychology, University of Salzburg, 5020 Salzburg, Austria; Michael.Liedlgruber@sbg.ac.at (M.L.); Frank.Wilhelm@sbg.ac.at (F.H.W.)
* Correspondence: guenther.sagl@gmail.com (G.S.); bernd.resch@sbg.ac.at (B.R.)

Received: 22 August 2019; Accepted: 10 October 2019; Published: 14 October 2019

Abstract: Wearable sensors are increasingly used in research, as well as for personal and private purposes. A variety of scientific studies are based on physiological measurements from such rather low-cost wearables. That said, how accurate are such measurements compared to measurements from well-calibrated, high-quality laboratory equipment used in psychological and medical research? The answer to this question, undoubtedly impacts the reliability of a study's results. In this paper, we demonstrate an approach to quantify the accuracy of low-cost wearables in comparison to high-quality laboratory sensors. We therefore developed a benchmark framework for physiological sensors that covers the entire workflow from sensor data acquisition to the computation and interpretation of diverse correlation and similarity metrics. We evaluated this framework based on a study with 18 participants. Each participant was equipped with one high-quality laboratory sensor and two wearables. These three sensors simultaneously measured the physiological parameters such as heart rate and galvanic skin response, while the participant was cycling on an ergometer following a predefined routine. The results of our benchmarking show that cardiovascular parameters (heart rate, inter-beat interval, heart rate variability) yield very high correlations and similarities. Measurement of galvanic skin response, which is a more delicate undertaking, resulted in lower, but still reasonable correlations and similarities. We conclude that the benchmarked wearables provide physiological measurements such as heart rate and inter-beat interval with an accuracy close to that of the professional high-end sensor, but the accuracy varies more for other parameters, such as galvanic skin response.

Keywords: wearable sensors; psychophysiology; sensor data analysis; time series analysis; signal analysis; similarity measures; correlation statistics; quantitative analysis; benchmarking

1. Introduction

In the last decade, the body of literature about physiological sensing and deriving emotions from physiological parameters has grown significantly. One reason for this is the rapid increase in variety of affordable wearable sensors that measure a broad range of physiological parameters such as heart rate, galvanic skin response, skin temperature, and others. With this increase, the "Quantified Self" community that promotes the idea of 24/7 tracking and monitoring has been growing significantly [1–3].

These new low-cost wearables are increasingly used in scientific studies in a variety of areas like health research, well-being assessment, disaster management, emotion information extraction and spatial emotion analysis, and stress detection [4–13]. However, some research efforts have used

wearable physiological sensors without prior investigation of the sensor's exact quality parameters, i.e., how accurately a sensor actually measures a given parameter or how reliable a sensor is in producing continuously high-quality measurement results.

Understanding a sensor's quality and accuracy is critical because the research results may otherwise be unreliable: while traditional professional wired sensor devices, which have been used for some time in laboratory and ambulatory studies in the fields of psychological and medical research, are proven to be highly accurate, most wearable sensors used in previous studies are not. In fact, most of them are not medically and/or electronically certified, which compromises the reliability of the measurement results. However, recently, some wearable sensors have been released that are certified and comply with a number of international standards (sensor technology, wireless communication, data transmission, etc.), which makes them a viable alternative to traditional wired equipment.

In the context of this research, we aim to investigate the measurement quality of two wearable sensor devices, namely the Zephyr BioHarness 3 and the Empatica E4, by comparing their measurements to those of calibrated laboratory sensors. Concretely, we are interested in the similarity and correlation of univariate time series from two different sensors that measure the same physiological parameters at the same time on the same participant. To evaluate the accuracy of the low-cost sensors, we perform benchmark testing between low-cost sensors against high-quality and well-calibrated sensors that act as the trusted gold standard. The second aim of this research is to detect and quantify relationships and dependencies between pairs of the same and different physiological parameters measured by different sensors. Our study assesses the parameters heart rate (HR), inter-beat interval (IBI), and galvanic skin response (GSR).

The remaining part of the paper is structured as follows. In Section 2, we provide a concise summary of related work regarding sensor benchmarking, followed by an overview of the physiological parameters of interest and the sensors used for this research (Section 3). The benchmarking methodology is presented in Section 4, where we also explain the entire workflow from sensor data acquisition to the analysis results. Section 5 descriptively illustrates the results, including a variety of statistical visualisations of similarity and correlation patterns. Finally, we discuss the results obtained and close the paper with our core conclusions.

2. Sensor Benchmark Methods—Related Work

The analysis of physiological signals from wearable sensors in order to better understand the human emotional response to the immediate surroundings has been investigated for several years. In recent years, a variety of affordable wearable sensors that measure well-established physiological parameters, such as heart rate and galvanic skin response, has reached the market. As a logical consequence—and as already mentioned in the introduction—the "Quantified Self" community is growing faster than ever, and inspiring scientific research, especially related to emotion and stress detection [5,7–9,14–17]. In any case, the basis for any further advanced analyses is adequate data quality in terms of accuracy, reliability, and validity [7,9,18]. However, scientific literature about the similarity and correlation of the measurements from such affordable wearables compared to those from well-calibrated and high-quality sensors from scientific laboratories is rare.

2.1. Similarity Measures

Generally speaking, the term 'similarity' is not rigorously mathematically defined. A variety of similarity measure families exist, for instance, distance-based (e.g., Euclidean distance), feature-based (e.g., Fourier coefficients), model-based (e.g., autoregressive), and elastic measures such as Dynamic Time Warping (DTW) and Edit Distance on Real sequence EDR [19–22]. A comprehensive review, however, is out of the scope of this paper—the interested reader may refer to [19,20,23,24], among other work.

In this research, we go beyond global measures and linear models to assess similarity. To uncover local similarity characteristics of time series, we thus follow a moving window approach combined

with more informative distance metrics. Elastic measures, such as DTW and the Fréchet distance, allow for a one-to-many comparison of time series elements, while so-called "Lock-Step" measures, such as Euclidian and Manhattan distance, only allow comparison of fixed pairs, making them very sensitive to local time-shifts and noise [23].

DTW temporally aligns two time series using the shortest path in a distance matrix, i.e., the path with the minimal global warping distance [25,26], thereby finding the most representative distance of the overall difference [20]. However, a comprehensive experimental comparison of representation methods and distance measures of time series reveals inconsistencies and even contradictions in the observations reported in individual studies [23]. An important consequence of this is that experimental results cannot be generalised without critically reviewing the assumptions made for a particular research context and study design. As concluded in [23], "there is no clear evidence that one similarity measure exists that is superior to others in the literature in terms of accuracy. While some similarity measures are more effective on certain data sets, they are usually inferior on some other data sets" (p. 297). The DTW distance outperforms Euclidian distance in a variety of studies [27]. Other types of measures are "Edit measures" and "Threshold measures". The former type includes, for instance, Longest Common Sub-Sequence LCSS, Edit Distance on Real sequence EDR and Edit Distance with Real Penalty ERP. The latter type includes Tightness of Lower Bounds TLB. The accuracy of the aforementioned other types is close to the accuracy of DTW, but DTW is much simpler [23,28]. We thus concluded to use DTW to assess the temporal similarity of the physiological time series.

To assess the geometric shape of a curve or curve segment, other distance measures, such as the Fréchet distance [29], can be used [30–32]. "The Fréchet distance is typically explained as the relationship between a person and a dog connected by a leash walking along the two curves and trying to keep the leash as short as possible. The maximum length the leash reaches is the value of the Fréchet distance" [33] (p. 7). We thus use the Fréchet distance to assess the geometric similarity of time series of sensor measurements of the same physiological parameter (e.g., GSR) on the same participant at the same time but with different sensors.

2.2. Correlation Statistics

The correlation of time series has been investigated for decades, in diverse fields. Herein, our focus on time series correlation is twofold: first, the correlation between equal-type physiological parameters measured by different sensors at the same time on the same participant in order to quantify differences between low-cost and un-calibrated sensors versus high-end and calibrated laboratory sensor equipment; second, the correlation between physiological parameters of different types, for instance, IBI and GSR, to explore potentially hidden relationships.

According to [34], the Pearson's correlation coefficient is the most robust metric when measuring the similarity in physiological time series—where robustness is understood as insensitivity to small variations. However, Pearson's r is highly sensitive to outliers and only considers linear relationships. Spearman's rank correlation coefficient (rho) is—as the name says—based on the rank of the values rather than on the values themselves; thus, it measures monotonicity rather than linearity. Therefore, using Spearman's rho to measure the strength of the associations between two variables leaves room for interpretation [35].

The human cardiovascular system and the autonomic nervous system are highly non-linear systems. In order to explore possible underlying non-linear interactions in the relationship between different physiological parameter, we herein, use the Maximum Information Coefficient (MIC) [36,37]. Several studies show the possibility of gaining new insights into such non-linear interactions when applying the MIC, for instance, in the interactions between neural and respiratory dynamics [38].

Further, one method to assess the temporal lag (or lead) between pairs of time series is the cross-correlation function in the time domain [39]. To get meaningful cross-correlation results, the time series need to be stationary, i.e., have a constant mean and variance. Time series stationarity

can be tested using, for instance, the Augmented Dickey-Fuller test [40]. Unless the time series is stationary, it needs to be differenced and tested for stationarity.

3. Physiological Parameter of Interest and Sensors used for Benchmarking

Herein, we describe the physiological parameters we investigated, and the sensors used to measure them. We investigated three sensors and four physiological parameters (Table 1):

- HR: heart rate, i.e., heartbeat frequency, unit: beats per minute
- IBI: inter beat interval, also known as the RR interval, i.e., the time between two R-peaks in the ECG's QRS complex, unit: milliseconds
- ECG: electrocardiogram, i.e., electrical activity of the heart, unit: millivolt
- GSR: galvanic skin response, i.e., the level of electric conductance of the skin, unit: microSiemens [µS]

Table 1. Benchmarked sensors and physiological parameters of interest.

	VarioPort	Zephyr BioHarness 3	Empatica e4
			
HR	X	X	–
IBI	X	X	X
ECG	X	X	–
GSR	X	–	X

3.1. VarioPort

The VarioPort (http://www.bisigma.de) is a small, lightweight, and highly flexible recording system that is used for multi-channel physiology recordings in laboratory and ambulatory setups. The standard version of the device can record up to 16 signals from connected pre-amplifiers (e.g., electromyography, electrocardiography, electrodermal activity, or respiration). The device has two built-in marker buttons that can be used to signal certain events occurring over time, resulting in an additional channel of data. We used these buttons to mark changes in the activity phases (for details refer to Section 4.1). Available sensors are either wire-connected to the device or are directly integrated into the device. The recorded data are stored on an SD card inside the device. The VarioPort allows setting different sampling rates for different channels, thus effectively reducing the required storage, especially in case of slowly changing signals (e.g., such as the skin conductance). For rapidly changing signals, such as ECG, sampling rates of up to 1024 Hz can be set. Since the VarioPort is the platform used for scientific studies at our Psychology Department, we used it as the gold standard in our benchmarking. In the remaining part of the paper, the VarioPort sensor is called VP.

3.2. Zephyr BioHarness 3

The Zephyr BioHarness 3 (https://www.zephyranywhere.com/) is a multivariable physiological monitoring device with a chest belt sensor that measures a wide variety of physiological parameters. The BH is a certified medical product (FDA Class II). Due to its design as a chest belt, the BioHarness 3 can measure ECG, RR intervals, respiration frequency, and other parameters such as 3D acceleration on a single sensor platform. Furthermore, parameters such as heart rate can be derived from the directly measured parameter, for instance, from ECG (HR within 0–240 BPM and an accuracy of ±1 BPM). The sampling rate for ECG is 250 Hz. The BH and the smartphone are connected wirelessly via Bluetooth. The raw data are accessible via a free SDK in binary format, and the device has been

extensively tested and validated in practical applications [41,42]. In the remaining part of the paper, the Zephyr BioHarness 3 sensor is called BH.

3.3. Empatica E4

The Empatica E4 (https://www.empatica.com/research/e4/) is a wrist band sensor that measures HR and GSR, as well as other parameters. The E4 is medically certified according to CE Medical 93/42/EEC Directive, class 2a, FCC. The sampling rate for GSR is 4Hz and for IBI 64Hz. According to Gradl et al. [18], the E4 is a wearable sensor that has the potential to measure mental stress. The E4 allows access to the raw data via smartphones through a comprehensive SDK and a Bluetooth connection. In the remaining part of the paper, the Empatica E4 sensor is called E4.

4. Benchmark Method

4.1. Study Setup and Participants

Our study included 18 participants who were recruited via e-mail. The test group comprised nine females and nine males in an age range of 24 to 40 years. All test persons were physically in decent shape and did not suffer from any illness at the time of the study.

The study was carried out at the University of Salzburg's Department of Psychology. After the study leaders attached the sensors, the participants were instructed to sit on an ergometer and follow the following routine:

- Resting phase: 5 min of rested seating on the ergometer, not performing any physical activity; used as a calibration phase for the measurements
- Cycling phase: 10 min of cycling at a constant 50 rpm; stepwise increase of physical load (5 steps of 2 min each, during which the resistance/power of the ergometer was increased from 35–65–100–133–165 W)
- Cool down phase: 5 min cool down, rested seating like in the resting phase

Participants were told not to interact with other people in the room, to focus on their task, and not to perform any physical activity other than as instructed. This was observed by the study leaders. For each 'run', two test persons were doing the lab study in parallel next to each other. Before commencing the actual exercise, we checked that all sensors were well positioned according to the participants' individual body shape. Additionally, we used surgical tape as necessary to hold the devices in place to make sure that we receive plausible measurements. We conducted these checks for each participant individually. All participants were aware of the aim of this research, and we obtained informed consent from all participants prior to commencement of the study.

4.2. Data Acquisition

The basic data acquisition workflow is illustrated in Figure 1. Each participant was equipped with diverse sensors to measure the physiological parameter using different platforms, namely VarioPort (VP), Zephyr BioHarness3 (BH), and Empatica E4 (E4). For each run, which refers to a participant exercising while their physiological parameters are measured, the raw sensor data are either stored in an SQLite database directly on the smartphone, or as files in a proprietary format on an SD card. The measurements from VP were extracted to flat files using the Software ANSLAB [43]. In contrast, the measurements from BH and E4 were sent to and fused by the e-Diary App into an SQLite database. The "raw sensor data" serves as input for the pre-processing, which is necessary to prepare the data for further analyses. The e-Diary App is herein purely used for sensor data collection and data management. During real-world field studies, however, the e-Diary App collects additional data such as GPS positions and contextual user feedback used for ground-truthing, thereby enabling the investigation of moments of stress in a spatio-temporal and contextual manner [4,44].

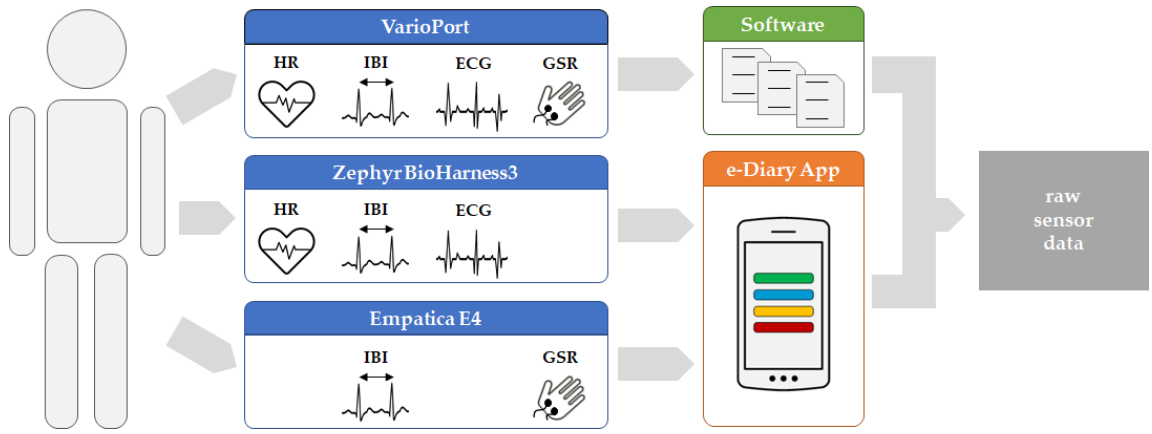

Figure 1. Data acquisition workflow—from the human participant (**left**) to raw sensor data (**right**).

4.3. Data Pre-Processing

The "raw sensor data" from the previous step serves as input for the data pre-processing phase, which is illustrated in Figure 2.

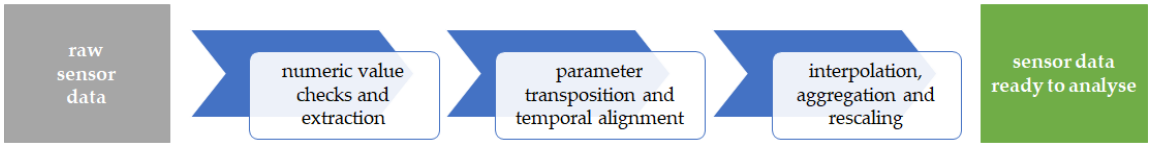

Figure 2. Data pre-processing phase—from raw sensor data (**left**) to sensor data ready to analyse (**right**).

This pre-processing phase consists of three main steps:

1. Extracting numeric values from differently encoded strings of values:

Data in the SQLite DB are stored in 1 s intervals in different formats due to various sampling rates of each of the sensors measuring different physiological parameters. For instance, the Empatica E4 measures GSR at a sampling rate of 4 Hz, while the VarioPort measures GSR at a sampling rate of 25 Hz.

The result is a table with sensor values where each single measurement has a correct timestamp.

2. Transposition of parameters and temporal alignment of measurements:

First, the vertical parameter structure (one row consists of a timestamp and a single sensor measurement, the next row consists of the same timestamp and with another single sensor measurement) needs to be transposed to a horizontal structure (a common timestamp and individual values as columns: one row consists of a timestamp and all sensor measurements that occurred at that timestamp).

Second, the irregular timestamps of all measurements are aligned to the millisecond in order to ensure the best possible time matching to the sensors' synchronized time. Since a 1 millisecond resolution is below the original sampling period, the measurements are aggregated depending on the parameters.

The result is a regular multivariate time series with 10 or 100 millisecond resolutions where some parameters at some timestamps may be missing values while other parameters are averaged within the given millisecond interval.

3. Interpolation, moving average and rescaling:

To fill missing values introduced by the temporal alignment in step 2, we applied spline interpolation, because it tends to greatly reduce oscillation by taking into account data points before and after the gap to be interpolated for a continuous representation [45]. In addition to the raw data, we calculated a moving averaged version with a window of ±5 s to eliminate high local variations.

For the correlation analysis of same type parameters and exploratory plots, we keep the original scaling of individual time series to identify potential offsets of measurements. For the similarity analysis, however, we rescale the measurements of individual time series from minimum and maximum to 0 and 1 in order to compare similarity distance metrics.

The data pre-processing was mainly carried out directly in the database using SQL and Java.

4.4. Statistical Signal Analysis—Time Series Correlation and Similarity Analysis

This sub-section illustrates how we assessed the correlations and similarities between time series of the same physiological parameters measured by different sensor platforms. Additionally, for some selected statistics, such as the MIC, we also run the analysis between different physiological parameters in order to explore potentially unknown relationships. The basic analytical workflow is shown in Figure 3. Note that we use the original signal scaling for exploratory plots, linear regression, and cross-correlation, while we use the rescaled signal (minimum → 0 . . . maximum → 1) to get similarity measures such as Fréchet and DTW distance.

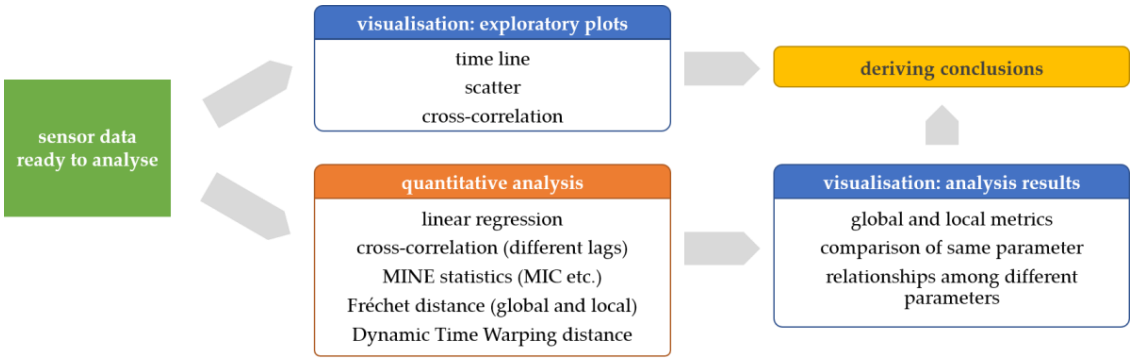

Figure 3. Data analysis workflow—from sensor data ready to analyse (**left**) to visualizations of exploratory plots and quantitative analysis results to deriving conclusions (**right**).

Of the complete time series derived from the pre-processing workflow (see Section 4.2), we focused on the following physiological parameters: HR, GSR, IBI, ECG. Additionally, from the ECG signal, we again derived the IBI and the Complex Demodulation amplitudes for the following frequency bands to estimate the heart rate variability [46]:

- Very low frequency VLF (0.025–0.07 Hz)
- Low frequency LF (0.07–0.14 Hz)
- High frequency HF (0.14–0.5 Hz)

In order to quantify pairwise correlations and similarities, we focused on:

- The linear regression coefficient of determination R^2, to assess the fit of pairwise timer series to a linear model [23]
- Cross-correlation, to assess temporal shifts [47]
- Maximal Information-based Nonparametric Exploration MINE statistics, in particular, the MIC, to assess functional associations, and MIC-R^2 to assess non-linear associations [36,48]
- Fréchet distance, to explore geometric similarity [30,49]
- DTW distance, to explore temporal similarity [26,28]

Signal analysis and most plots were done using the statistical computing software R, while some other plots were produced with the data visualisation software Tableau Desktop.

5. Results

For each of the 18 participants, we captured 15 parameters either measured directly or derived from ECG. For all these parameters, we also computed a moving average for an outlier-smoothed version of the same signal in order to get a better understanding of the signal's overall robustness and reliability. Additionally, we computed low/high-pass filtered versions of the GSR signals, as well as the complex demodulation amplitudes from the ECG signals. For each participant, we analysed 22 pairs of physiological parameters of the same type regarding similarity (e.g., heart rate from BioHarness sensor and heart rate from VarioPort sensor), and another 136 pairs of parameters of different type regarding correlations (e.g., heart rate from BioHarness sensor and galvanic skin response from VarioPort sensor).

Since there are many different parameters, we defined a naming convention that includes the physiological parameter of interest, the platform used to measure it, plus an indication of whether a time series is a moving averaged and/or a filtered version. For the naming of these parameters, we use the following notation:

For direct measurements:

<*parameter*> <*platform*> [*filt.*] [(*mv. avg*)]

where *parameter* can be GSR, HR, or IBI and *platform* can be BH, E4, or VP; *(mv. avg.)* indicates that this is the moving averaged version; *filt.* indicates that a first order high-pass (0.05 Hz) and first order low-pass (0.5 Hz) Butterworth filter has been applied to the original signal (this filter setting is used for further analysis to identify moments of stress [4,11]; however, this is not within the scope of this paper).

example: *GSR: VP (mv. avg.)* refers to the moving averaged version of galvanic skin response measured by VarioPort

For derived measurements:

<derived parameter> from <original parameter> <platform> [(mv. avg.)]

where *derived parameter* can be HF, IBI, LF, VLF, *original parameter* can be ECG, and *platform* can be BH, or VP; *(mv. avg.)* indicates that this is the moving averaged version example: *IBI: from ECG BH* refers to the inter beat interval derived from the electrocardiogram measured by BioHarness

The following subsections are structured according to Figure 3. We use two representative time series, one HR and one GSR, as examples to guide the reader through the high number of physiological parameters investigated herein. These two examples are cross-referenced between several figures and thus provide views on the same data from different perspectives, thereby fostering the consolidation of a more holistic picture.

5.1. Visualisation: Exploratory Plots

The aim of the exploratory plots is to obtain a basic understanding of the temporal behaviour and the relationship of equal-type physiological parameters. For this first insight, we investigate three complementary plots that provide different views on the same data: a time series plot, a scatter plot, and a cross-correlation plot. These plots show two versions of the same parameter, namely a data-as-is version and moving averaged version. Note that for the cross-correlation plots the second parameter is used as the independent one. To illustrate the methodology and exemplary results, we only show representative sample plots, which highlight the characteristic patterns of about 80% of all plots. Overall, we produced more than 1000 plots based on unscaled and rescaled data.

Figures 4 and 5 each show a time plot (a), a scatter plot (b), and a cross-correlation plot (c,d). The physiological parameter of interest is HR, measured by HB and VP. The HR time plot (Figure 4a) shows two highly similar, almost identical curves. The blue curve has an offset, which is maximal in the low range and converges to zero in the high range. The corresponding error term seems

to include a reciprocal component: the higher the actual measurement, the lower the error. The HR scatter plot (Figure 4b) shows a high positive linear relationship with an R^2 of 0.971 for raw data and 0.997 for the moving averaged data. This means that 97.1% and 99.7%, respectively, of the data's total variance can be explained by a linear model. This plot also confirms what is seen in the time plot, namely that in the low range the residuals are higher than in the high range. Note that the higher residuals in red in the upper right quarter of the plot refer to the time plot at ~750 s, where the blue curve drops below the red curve (indicated by a black arrow in Figure 4a,b). The HR cross-correlation plots show the highest cross-correlation for the as-is version (Figure 4c) at a lag of 1 s, and the highest cross-correlation for the moving averaged version (Figure 4d) at a lag of 2 s. In other words, the local trend of the BH is lagging 1 and 2 s, respectively, "behind" the local trend of the VP on average.

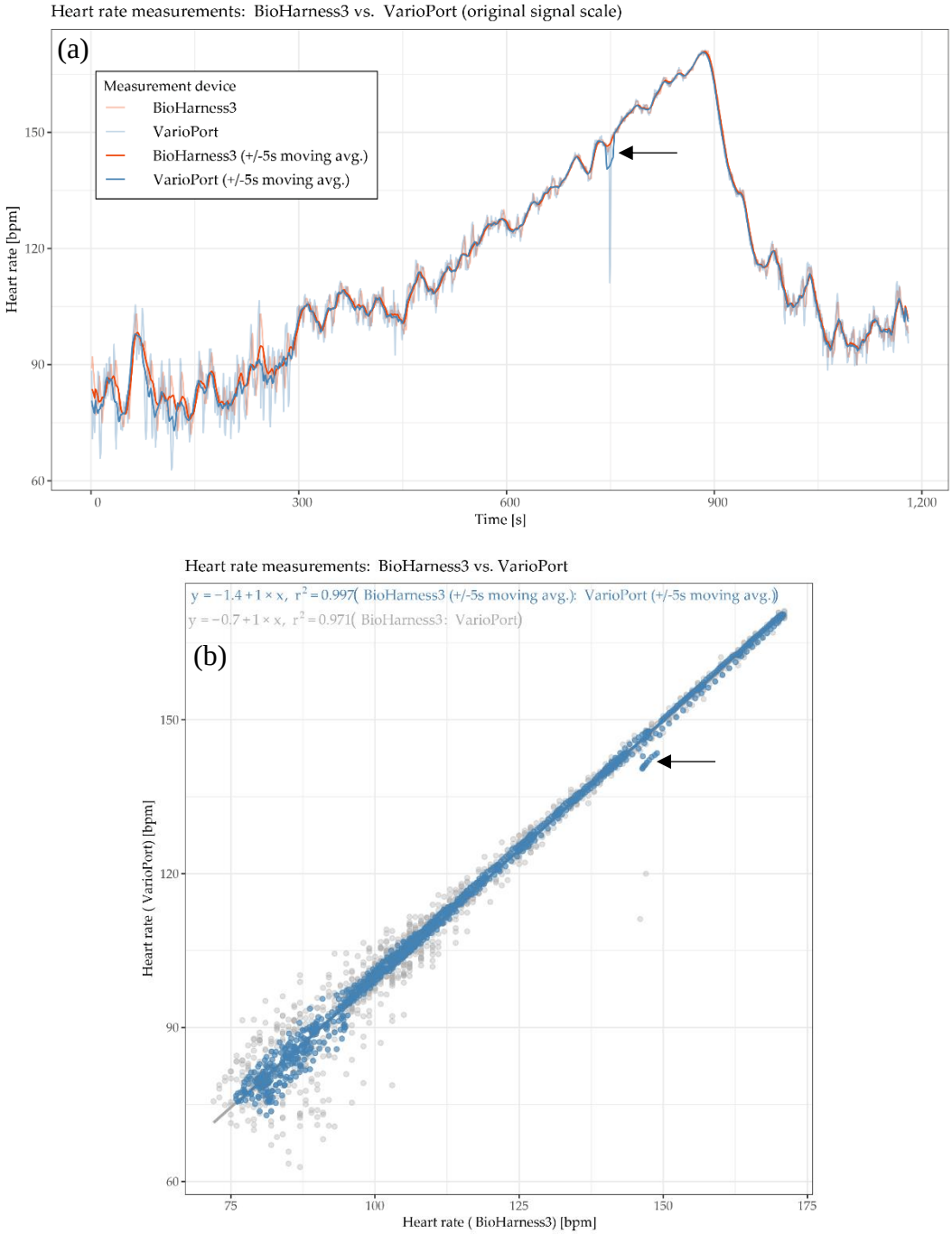

Figure 4. *Cont.*

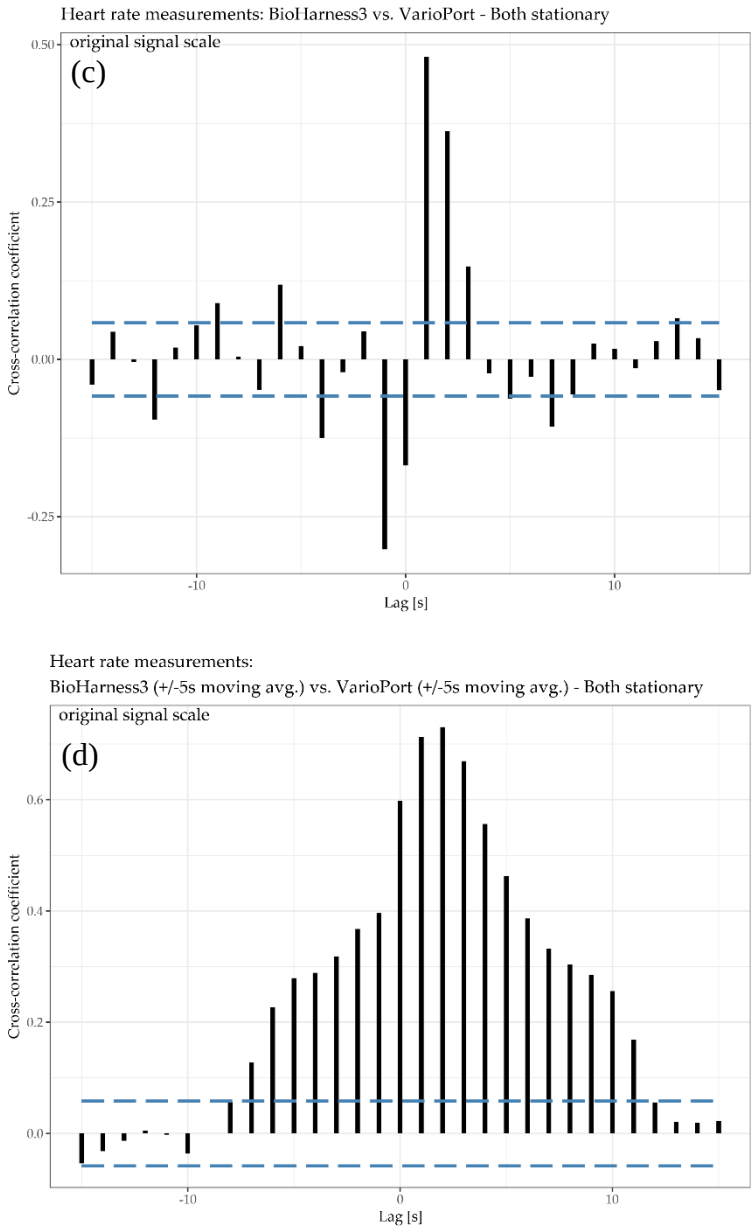

Figure 4. Participant RP 5–14: time plot (**a**), scatter plot (**b**), and cross-correlation plot (**c**), and cross-correlation plot of moving averages (**d**) of heart rate HR [beats per minute] measured by Bioharness3 BH sensor and VarioPort VP.

In Figure 5, the physiological parameter of interest is GSR, measured by the E4 and the VP. Generally speaking, measuring GSR is, in comparison to HR, a delicate undertaking due to the measurement principle, which is solely based on the electrical conductivity of the skin. This conductivity is highly dependent on (1) the participant's skin characteristics and (2) the contact between the sensor electrodes and the skin, especially during physical activity. Thus, these two factors can have a significant impact on the reliability, and thus on the comparability, of the sensor measurements. Additionally, the mounting of the sensors' electrodes can differ as well. For instance, the VP electrodes need an isotonic electrolyte gel to ensure reliable measurements, while E4 does not require anything. The GSR time plot (Figure 5a) shows that the blue curve (VP) increases gradually. The red curve (E4) increases faster than the blue one and shows a local maximum after the 5-min warm-up phase (at around 300 s). This increase is followed by a decrease for another 5 min (until around 600 s). Aside from a little drop at around 900 s, the red curve increases until the end of the cool down phase. In general, the E4 seems to be

more responsive to sweating associated with physical effort than the VP, which may be due to its lack of stabilizing isotonic electrolyte gel. The GSR scatter plot (Figure 5b) shows a positive correlation with an R^2 of 0.882 for raw data and 0.896 for the moving averaged data. This means that 88.2% and 89.6%, respectively, of the data's total variance can be explained by a linear model. The cross-correlation plots (Figure 5c) show the highest cross-correlation for the as-is version at lag of 2 s, and the highest cross-correlation for the moving averaged version at a lag of 1 s. In other words, the local trend of the E4 sensor is lagging 2 and 1 s, respectively, "behind" the local trend of the VP sensor on average.

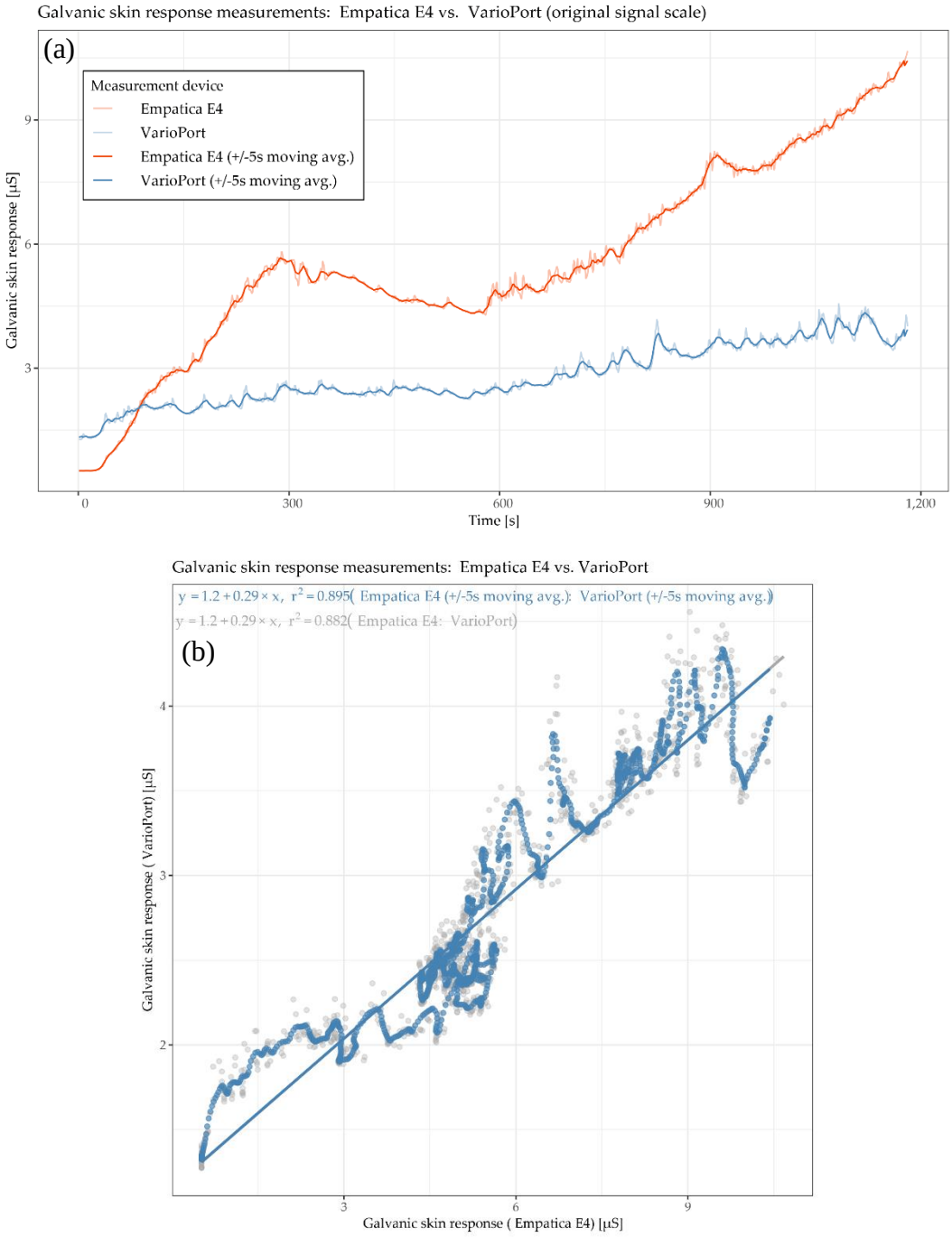

Figure 5. *Cont.*

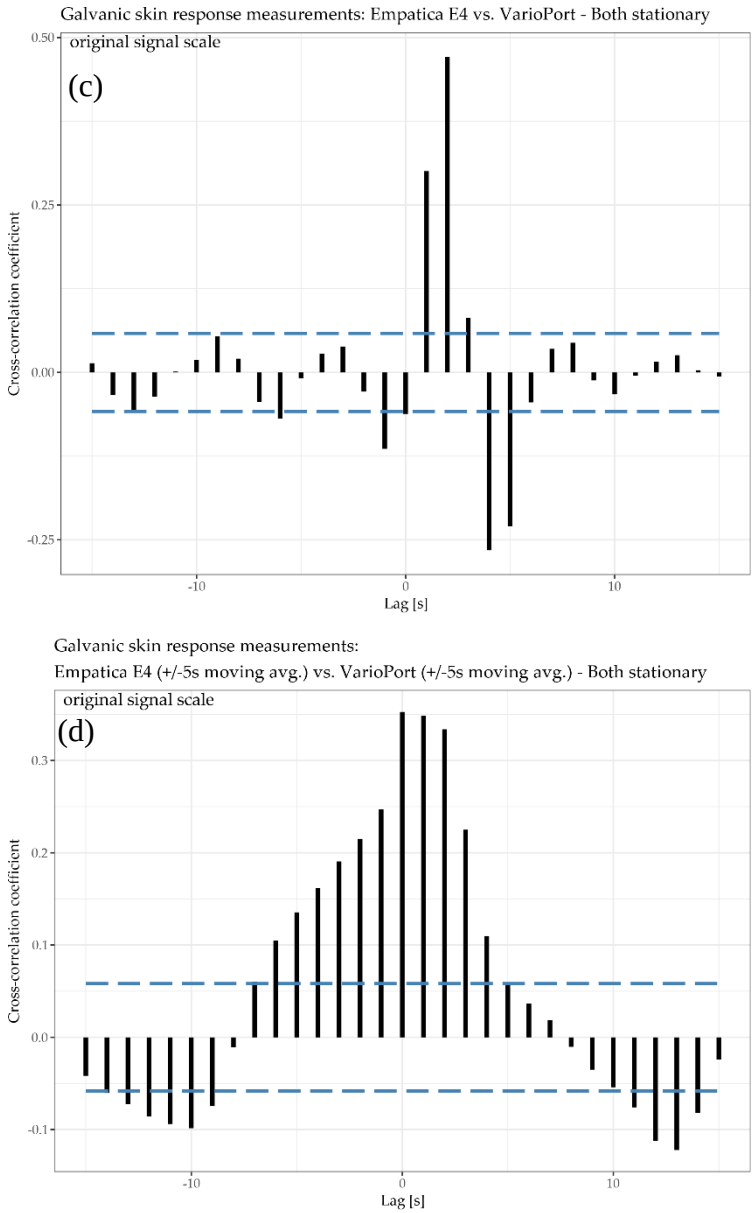

Figure 5. Participant RP 3–8: time plots (**a**), scatter plot (**b**), cross-correlation plot (**c**), and cross-correlation plot of moving averages (**d**) of galvanic skin response GSR measured by Empatica E4 and VarioPort.

5.2. Quantitative Analysis

The aim of the quantitative analysis is twofold: first, we assess the correlation and the similarity of equal-type parameters. Second, we assess potential associations in pairs of parameters of both equal-type and different types. In addition to global statistics, we also apply local measures to derive new information about the relationship between and among the different parameter pairs. This combination of global and local similarity and correlation measures on the individual level further enables a roll-up view on relationship patterns of physiological parameters among participants. Note that for similarity distance metrics, we rescaled measurements from min … max of the original scale to 0 … 1. For the correlation analyses we used the original values of the given parameter at the given range and the given unit in order to identify potential offsets.

5.2.1. Linear Regression and Coefficient of Determination R^2

The first statistic of interest is the coefficient of determination R^2, which quantifies the percentage of the variance of the two given parameters that can be explained by a linear regression model. In addition to the R^2 of individual pairs of parameters, as shown in the scatter-plots (refer to Section 5.1, Figures 4b and 5b), we now investigate all pairs among all participants and explore the corresponding R^2 pattern. This pattern can be derived from the R^2 matrix shown in Figure 6. Furthermore, for each group of parameters, e.g., all IBI related parameter, we calculate the total average per participant in order to get an impression of the impact of each participant's individual overall measured activity (GSR base level, skin contact of electrodes, etc.).

Group	Parameter 1 - Parameter 2	RP 9-17	RP 9-18	RP 6-7	RP4-11	RP 2-6	RP 2-5	RP 5-14	RP 3-8	RP 4-12	RP 8-19	RP 7-19	RP 3-3	RP 6-21	RP 7-20	RP 5-13	RP 1-1	RP 8-20	RP 1-2	Total AVG
IBI	VP (mv. avg.) vs. IBI from ECG VP (mv.avg.)	1.00	1.00	0.98	1.00	1.00	1.00	1.00	0.88	0.99	1.00	1.00	0.66	1.00	1.00	0.56	0.99	1.00	1.00	0.95
	VP vs. IBI from ECG VP	1.00	0.99	0.94	1.00	0.98	1.00	0.99	0.70	0.91	0.98	0.99	0.30	1.00	0.99	0.30	0.96	0.99	0.98	0.89
	IBI from ECG BH vs. IBI from ECG VP	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.69	0.99	0.97	0.99	0.96	0.04	0.01	0.87
	BH (mv. avg.) vs. VP (mv. avg.)	0.99	0.96	0.99	0.95	0.99	0.97	0.99	0.87	0.94	0.89	0.97	0.95	0.73	0.54	0.53	0.38	0.76	0.01	0.80
	BH (mv. avg.) vs. IBI from ECG BH (mv. avg.)	0.99	0.96	0.98	0.96	0.99	0.98	0.99	0.99	0.95	0.90	0.98	0.95	0.74	0.54	0.95	0.45	0.01	0.01	0.79
	IBI from ECG HB vs. IBI from ECG VP	0.99	0.99	0.94	0.89	0.98	0.98	0.95	0.94	0.98	0.79	0.81	0.36	0.95	0.78	0.92	0.85	0.02	0.01	0.78
	BH vs IBI from ECG BH	0.71	0.42	0.84	0.68	0.67	0.38	0.79	0.61	0.55	0.28	0.58	0.42	0.30	0.17	0.55	0.14	0.00	0.00	0.45
	BH vs. VP	0.70	0.44	0.84	0.60	0.67	0.37	0.76	0.41	0.49	0.25	0.49	0.44	0.28	0.18	0.14	0.13	0.34	0.00	0.42
	E4 (mv. avg.) vs. VP (mv. avg.)	0.93	0.90	0.26	0.91	0.09	0.04	0.00	0.87	0.02	0.52	0.06	0.86	0.00	0.09	0.10	0.17	0.26	0.02	0.34
	E4 vs. VP	0.81	0.75	0.24	0.78	0.08	0.04	0.00	0.63	0.02	0.27	0.05	0.69	0.00	0.07	0.04	0.15	0.20	0.02	0.27
	Total AVG	0.91	0.84	0.80	0.88	0.74	0.67	0.75	0.79	0.68	0.69	0.69	0.63	0.60	0.53	0.51	0.52	0.36	0.21	0.66
HR	BH (mv. avg.) vs. VP (mv. avg.)	1.00	0.98	1.00	0.99	1.00	0.99	1.00	0.96	0.99	0.94	0.98	0.97	0.98	0.97	0.71	0.52	0.97	0.32	0.90
	BH vs. VP	0.98	0.96	0.97	0.95	0.98	0.94	0.97	0.86	0.90	0.51	0.84	0.86	0.93	0.88	0.40	0.50	0.88	0.30	0.81
	Total AVG	0.99	0.97	0.98	0.97	0.99	0.97	0.98	0.91	0.94	0.72	0.91	0.91	0.96	0.93	0.56	0.51	0.92	0.31	0.86
GSR	E4 (mv. avg.) vs. VP (mv. avg.)	0.22	0.73	0.77	0.92	0.49	0.59	0.66	0.90	0.80	0.09	0.09	0.89	0.06	0.95	0.73	0.57	0.50	0.00	0.55
	E4 vs. VP	0.21	0.72	0.77	0.91	0.49	0.58	0.65	0.88	0.79	0.05	0.09	0.89	0.06	0.93	0.73	0.57	0.49	0.00	0.55
	E4 filtered (mv. avg.) vs. VP filtered (mv. avg.)	0.91	0.67	0.42	0.02	0.33	0.68	0.09	0.13	0.23	0.92	0.56	0.06	0.18	0.28	0.28	0.16	0.54	0.49	0.39
	E4 filtered vs. VP filtered	0.90	0.60	0.41	0.01	0.37	0.64	0.10	0.13	0.22	0.87	0.49	0.05	0.21	0.23	0.30	0.14	0.54	0.38	0.37
	Total AVG	0.56	0.68	0.60	0.47	0.42	0.62	0.37	0.51	0.51	0.48	0.31	0.48	0.13	0.59	0.51	0.36	0.51	0.22	0.46
VLF ECG	BH (mv. avg.) vs. VP (mv. avg.)	0.90	0.95	0.99	0.96	0.99	0.99	0.99	0.89	0.99	0.78	0.97	0.88	0.97	0.66	0.81	0.29	0.05	0.01	0.78
	BH vs. VP	0.89	0.94	0.98	0.82	0.97	0.96	0.96	0.81	0.96	0.78	0.89	0.82	0.94	0.58	0.75	0.29	0.04	0.00	0.74
	Total AVG	0.89	0.95	0.98	0.89	0.98	0.97	0.98	0.85	0.98	0.78	0.93	0.85	0.96	0.62	0.78	0.29	0.04	0.01	0.76
LF ECG	BH (mv. avg.) vs. VP (mv. avg.)	1.00	1.00	0.99	0.96	1.00	0.99	0.99	0.96	0.99	0.98	0.98	0.90	0.98	0.59	0.77	0.22	0.01	0.00	0.79
	BH vs. VP	0.99	0.99	0.96	0.88	1.00	0.97	0.96	0.81	0.95	0.96	0.93	0.86	0.94	0.52	0.72	0.20	0.01	0.00	0.76
	Total AVG	0.99	0.99	0.97	0.92	1.00	0.98	0.98	0.89	0.97	0.97	0.95	0.88	0.96	0.55	0.74	0.21	0.01	0.00	0.78
HF ECG	BH (mv. avg.) vs. VP (mv. avg.)	0.98	0.99	0.99	0.96	1.00	0.98	0.99	0.95	0.99	0.99	0.95	0.63	0.95	0.05	0.02	0.29	0.01	0.03	0.71
	BH vs VP	0.89	0.92	0.81	0.56	0.95	0.79	0.90	0.52	0.81	0.91	0.58	0.34	0.71	0.04	0.08	0.06	0.00	0.00	0.55
	Total AVG	0.94	0.96	0.90	0.76	0.98	0.89	0.94	0.74	0.90	0.95	0.77	0.48	0.83	0.04	0.05	0.17	0.00	0.02	0.63

Figure 6. R^2 matrix of pairs of parameter and participants; detail (**a**) complements Figure 4, detail (**b**) complements Figure 5; total average of individual pairs among participants is shown in the last column; total average of participants among individual parameter pairs is shown in the last row of each parameter group; colour: red indicates high correlation, blue indicates low correlation.

In the upper half of the R^2 matrix, the pairs of equal-type parameters, measured by different sensors (or derived from another signal of the same sensor) show a high linear relationship across the majority of the participants. This relationship also indicates that these parameters are rather robust from a measuring point of view. However, the matrix also shows some cases with no relationship at all, see for instance IBI derived from ECG BH (moving averaged version), and IBI derived from ECG VP (moving averaged version) at row three for participant RP 8-20 and RP 1-2. This may indicate that one of the sensors did not have proper contact between the electrodes and the skin and thus failed to collect valid data.

The matrix shows that GSR in general, and IBI measured by Empatica E4, tend to have rather low or even no correlation, while some participants demonstrate the exact opposite (compare instance RP 4-11 and RP 2-5).

Note that the R^2 matrix in Figure 6 is organized as follows: for each group of parameters, the top row shows the highest correlation among all participants, while the bottom row shows the lowest correlation. Further, the left column shows the participant with the highest correlations among all parameters, while the right column shows the participant with the lowest correlations among all parameters.

Figure 6 detail (a) refers to the HR example shown in Figure 4 and detail (b) refer to the GSR example shown in Figure 5.

In summary, the overall pattern shown in Figure 6 confirms that both the type of the parameter measured and the individual parameters, such as skin characteristics of the participant, significantly influence the reliability of the measurements and thus the quality of further analysis.

5.2.2. Cross-Correlation

In order to assess potential temporal shifts between the measurements of the same physiological parameter, we investigate the cross-correlation at different lags. The corresponding cross-correlation pattern in Figure 7 shows that some pairs of parameters (especially the top five rows of the IBI group) have a rather low variance among the lags and tend to correlate positively (i.e., the lags within a cell show rather homogeneous coefficients). Other pairs of parameters (especially the lower half of the IBI group) have a rather high variance among the lags, indicating positive correlations around lag 0 and negative correlations at lag +15 and −15, respectively.

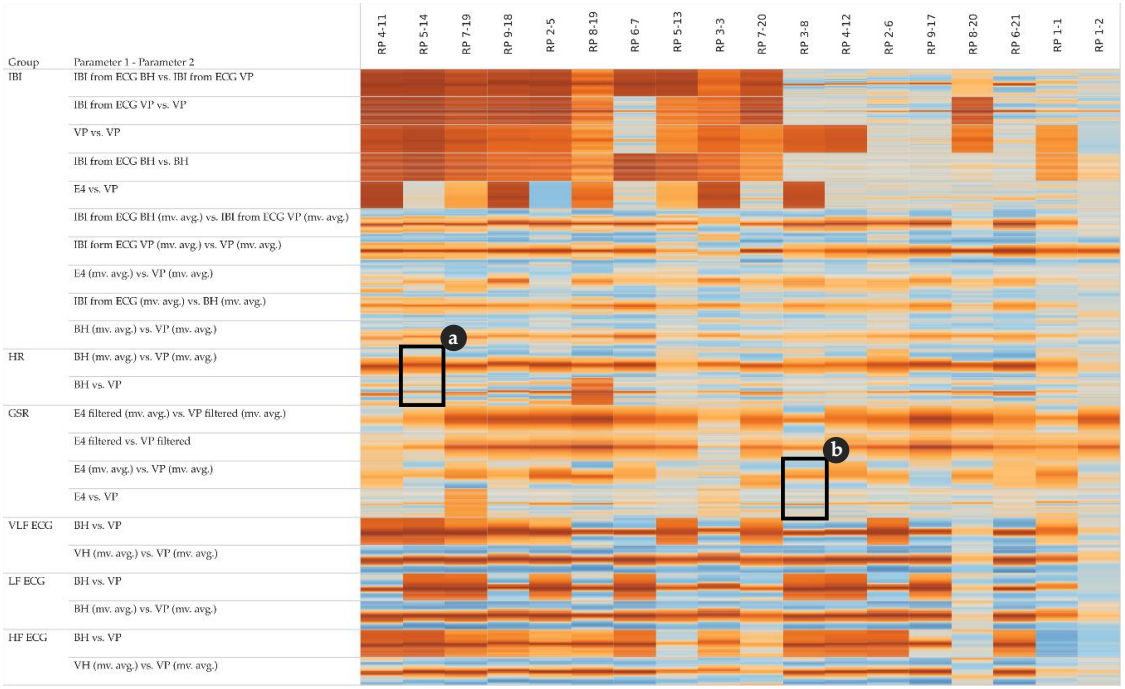

Figure 7. Cross-correlation matrix of pairs of parameters and participants; detail (**a**) complements Figure 4, detail (**b**) complements Figure 5; colour: orange indicates positive cross-correlation, blue indicates negative cross-correlation; a cell detail shows lags as small horizontal bars: lag −15 at top, and lag +15 at bottom.

Note that the cross-correlation matrix in Figure 7 is organized as follows: for each group of parameters, the top row shows the highest cross-correlations (i.e., lowest variance) among all participants, while the bottom row shows the lowest cross-correlation (i.e., highest variance). Further, the left column shows the participant with the highest cross-correlations among all parameters, while the right column shows the participant with the lowest cross-correlations among all parameters.

Figure 7 detail (a) refers to the HR example shown in Figure 4 and detail (b) refers to the GSR example shown in Figure 5.

5.2.3. MINE Statistics

By using MINE statistics, we investigate relationships between and among parameter pairs of same as well as different type, for instance, GSR versus HR. In addition to the linear correlation coefficient of determination R^2 (see Section 5.2.1), we use the Maximal Information Coefficient MIC to identify all functional relationships, also including linear ones as issued by R^2. Although a functional

relationship between certain combinations of parameters might be obvious (e.g., IBI [ms] = 60,000/HR [beats per hour]), we nonetheless include such combinations herein for reasons of confirmation.

Figure 8 shows 3 k-means clusters of pairs of different parameter types. We tested with k [1,5] and chose 3 because the result is most intuitive—the clusters show low, moderate, and high correlations. Figure 8a–f highlights some particularly interesting parts of the clusters:

- a, b: LF and VLF derived from ECG measured by BioHarness show a rather low correlation with almost all other parameters
- c, d: IBI measurements, either directly measured or derived from ECG, show a rather high correlation with HR and with IBI from other sensors
- e, f: GSR measurements from VP (high- and low-pass filtered) show a rather low correlation with all other parameters; however, GSR measurements from E4 (high- and low-pass filtered) show a low to moderate correlation with all other parameters.

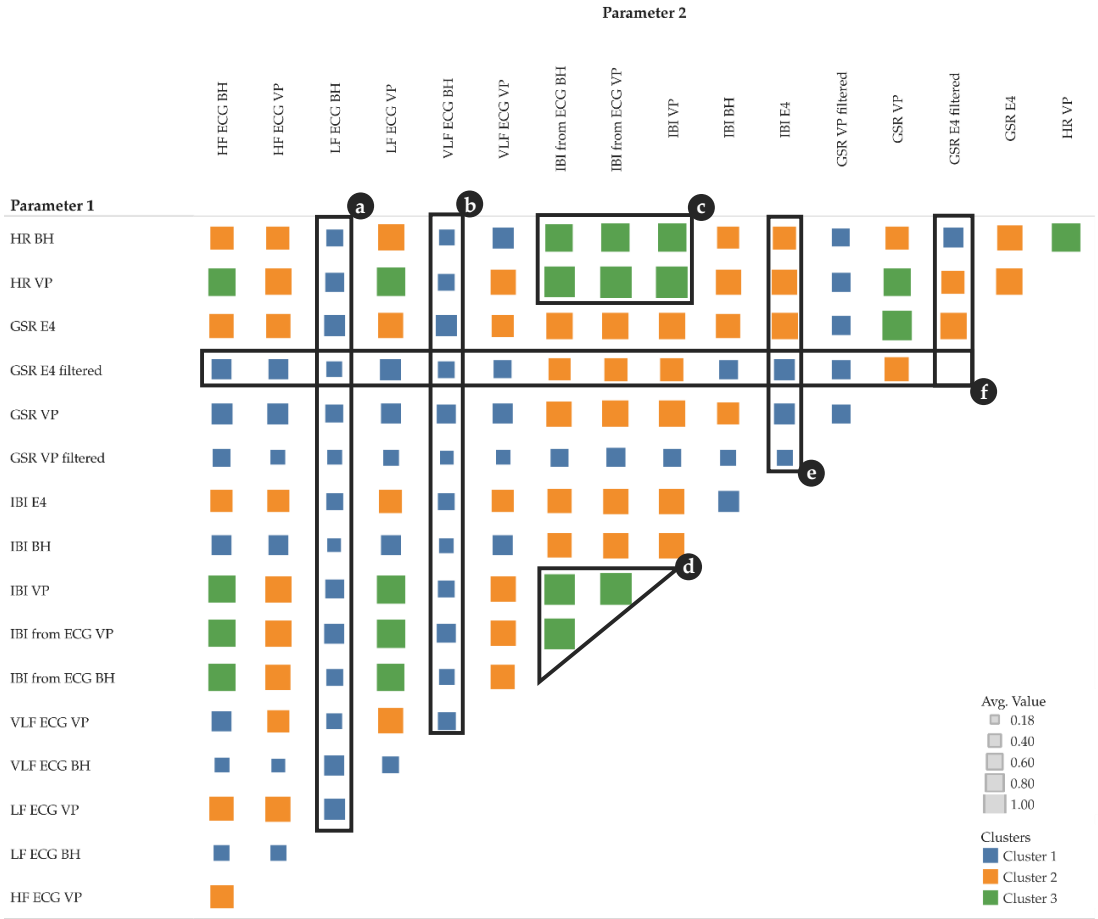

Figure 8. Maximum Information Coefficient (MIC) cluster matrix of pairs of parameters (size show averages among all participants, moving averaged versions only); colors: blue (cluster 1): low correlations; orange (cluster 2): moderate correlations; green (cluster 3): high correlations; symbol size in a matrix cell: average MIC among participants; (**a**–**f**) highlight special characteristics described in the text.

When focusing on the level of individual participants, Figure 9 shows MIC correlations among pairs of different parameters. Within a single cell, the small vertical bars represent participants (one bar per participant). Figure 9 complements Figure 8 by adding participant information to the corresponding clusters.

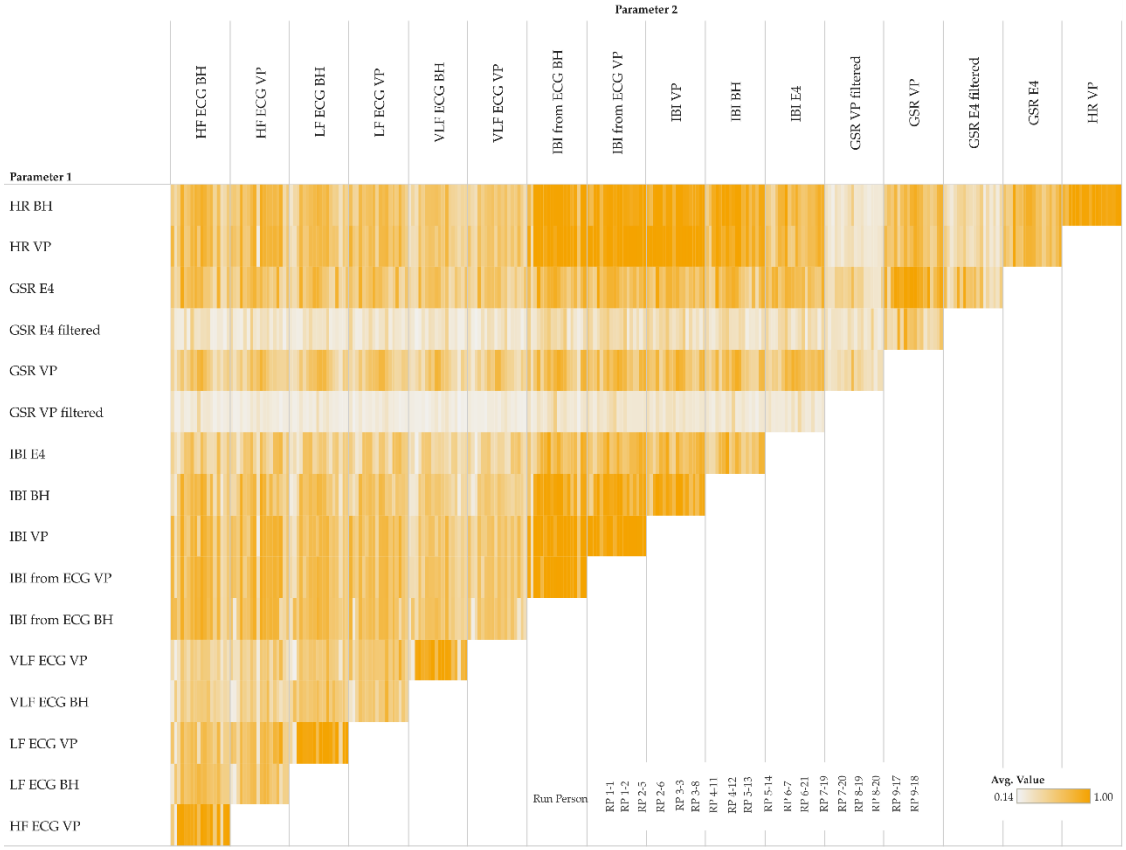

Figure 9. MIC participant-level matrix of pairs of parameters (moving averages only); details of a matrix cell show participants as small vertical bars (order of participants is shown in the legend).

The MIC is used to quantify the strength of any functional relationships, i.e., including linear ones, while the R^2 coefficient can only quantify linear relationships. By subtracting R^2 from the MIC, we compute a measure of nonlinearity [36], which we use to identify the following three classes of relationships as shown in Figure 10:

- "false" linear relationships (R^2 is not confirmed by MIC): MIC–$R^2 < 0$
- "true" linear relationships (R^2 is confirmed by MIC): MIC–$R^2 \sim 0$
- functional but not linear relationships: MIC–$R^2 > 0$

On the individual level, the MIC–R^2 matrix shown in Figure 11 provides additional detail to the clustering view. Interestingly, GSR measurements from E4 and VP show rather strong functional but not linear relationships with almost all other parameters (see third and fifth row in Figures 10 and 11). Particularly interesting is the relationship between GSR VP (filtered and moving averaged version, fourth row) and GSR E4 (filtered and moving averaged version, fifth-last column), which shows some highly negative values (see black arrow). These cases indicate a "false" linear relationship. For instance, participant RP 9–17: MIC 0.2 minus R^2 0.91 results in −0.71. In other words, the MIC does not confirm the highly linear relationship indicated by R^2; in fact, the MIC indicates that there is almost no relationship. From a physiological point of view, this relationship might be obvious; however, the quantification of this relationship from a data-driven perspective is, to our best knowledge, novel.

Figure 10. MIC–R^2 cluster matrix of pairs of parameters (moving averages only); colors: green (cluster 1): "false" linear relationships; yellow (cluster 2): "true" linear relationships; red (cluster 3): functional but not linear relationships.

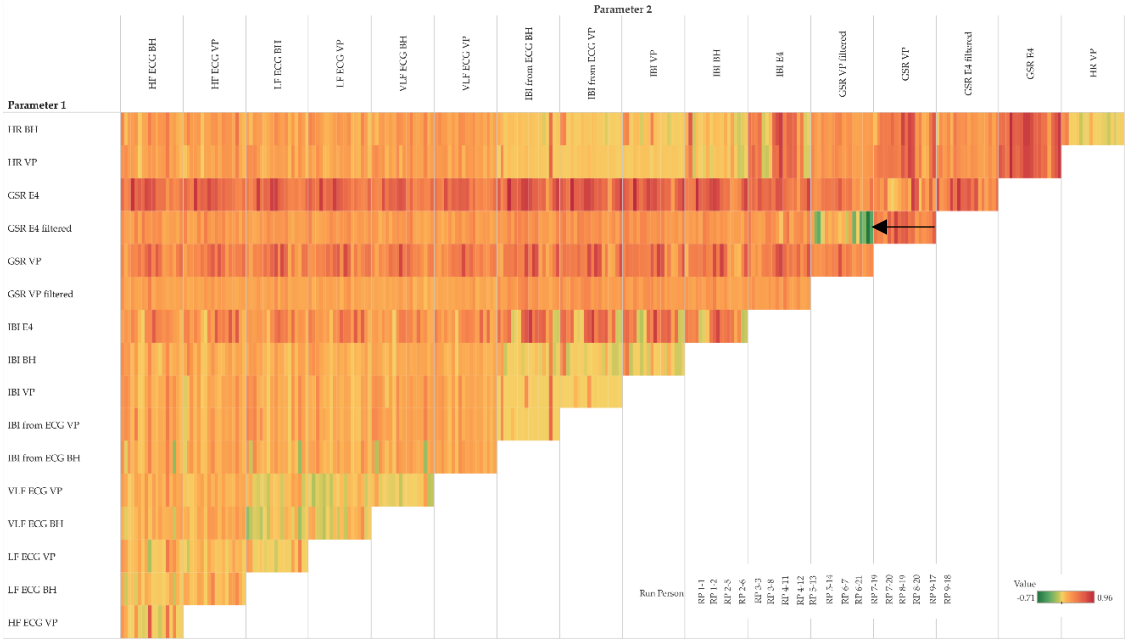

Figure 11. MIC–R^2 individual matrix of pairs of parameters (moving averages only); details of a matrix cell show participants as small vertical bars (order of participants is shown in the legend); back arrow points to pairs of parameters with very weak association.

5.2.4. Fréchet Distance (Global and Local)

The Fréchet distance is a measure of how different two curves are from each other in terms of geometric structure [32]. Herein we use the Fréchet distance to measure the geometric similarity of two time series of a physiological parameter measured by different sensor platforms, one being

professional and well-calibrated while the other is low-cost and wearable. In addition to the standard global Fréchet distance, we also compute local versions using a moving window approach.

The global Fréchet distance matrix (Figure 12) shows two expectable general aspects. First, cardiac parameters such as IBI, HF, LF and VLF derived from ECG seem to be more similar than GSR. This is likely because, from a measuring point of view, ECG-related measurements are simply more robust than, for instance, GSR-related ones. Second, the moving averaged versions of the time series also tend to be more similar than their non-averaged counterparts, which include more local fluctuations. Figure 12 detail (a) refers to the HR example shown in Figure 4 and detail (b) refers to the GSR example shown in Figure 5. In both detail (a) and detail (b), the moving averaged time series causes a smoothing effect, thus indicating a higher similarity as compared to the original (non-smoothed) time series.

Group	Parameter 1 - Parameter 2	RP 9-18	RP 9-17	RP 4-11	RP 6-7	RP 2-5	RP 5-14	RP 7-19	RP 4-12	RP 3-8	RP 2-6	RP 6-21	RP 8-19	RP 3-3	RP 7-20	RP 8-20	RP 5-13	RP 1-1	RP 1-2	Total AVG
IBI	IBI from ECG VP (mv. avg.) vs. VP (mv. avg.)	0.05	0.04	0.02	0.49	0.01	0.05	0.07	0.32	0.59	0.34	0.02	0.01	0.28	0.05	0.04	0.47	0.14	0.10	0.17
	IBI from ECG BH (mv. avg.) vs. IBI from ECG VP (mv. avg.)	0.03	0.04	0.09	0.05	0.05	0.04	0.12	0.05	0.08	0.08	0.08	0.51	0.36	0.36	0.45	0.19	0.26	0.86	0.21
	BH (mv. avg.) vs. VP (mv. avg.)	0.40	0.30	0.31	0.23	0.47	0.22	0.30	0.23	0.27	0.37	0.35	0.54	0.44	0.39	0.42	0.29	0.49	0.73	0.38
	IBI from ECG BH (mv. avg.) vs. BH (mv. avg.)	0.40	0.27	0.28	0.38	0.47	0.20	0.30	0.40	0.42	0.40	0.33	0.25	0.37	0.40	0.44	0.32	0.35	0.90	0.38
	IBI from ECG VP vs. VP	0.33	0.28	0.08	0.49	0.12	0.32	0.24	0.50	0.63	0.55	0.04	0.08	0.58	0.19	0.19	0.77	0.83	0.71	0.39
	ibi_from_ecg_bh - ibi_from_ecg_vp	0.12	0.13	0.24	0.10	0.15	0.16	0.23	0.11	0.20	0.45	0.13	0.55	0.55	0.50	0.99	0.65	0.88	0.89	0.39
	E4 (mv. avg.) vs. VP (mv. avg.)	0.30	0.31	0.34	0.89	0.61	0.90	0.75	0.80	0.59	0.73	0.77	0.60	0.60	0.68	0.73	0.81	0.85	0.70	0.67
	BH vs. VP	0.73	0.73	0.53	0.34	0.84	0.49	0.55	0.67	0.68	0.70	0.99	0.72	0.65	0.93	0.81	0.80	0.85	0.76	0.71
	IBI from ECG BH vs. BH	0.73	0.70	0.53	0.48	0.87	0.49	0.57	0.66	0.73	0.72	0.99	0.82	0.55	0.84	0.91	0.86	0.67	0.97	0.73
	E4 vs. VP	0.52	0.44	0.52	0.92	0.72	0.94	0.98	0.82	0.77	0.79	0.81	0.64	0.74	0.77	0.82	0.86	0.87	0.77	0.76
	Total AVG	0.36	0.32	0.30	0.44	0.43	0.38	0.41	0.46	0.50	0.51	0.45	0.47	0.51	0.51	0.58	0.60	0.62	0.74	0.48
HR	BH (mv. avg.) vs. VP (mv. avg.)	0.32	0.19	0.15	0.11	0.18	0.12	0.22	0.30	0.30	0.07	0.16	0.57	0.34	0.20	0.24	0.52	0.69	0.82	0.31
	BH vs. VP	0.50	0.69	0.27	0.30	0.45	0.30	0.42	0.67	0.42	0.60	0.31	0.67	0.49	0.52	0.35	0.90	0.71	0.98	0.53
	Total AVG	0.41	0.44	0.21	0.21	0.32	0.21	0.32	0.48	0.36	0.34	0.23	0.62	0.42	0.36	0.30	0.71	0.70	0.90	0.42
GSR	E4 filtered (mv. avg.) vs. VP filtered (mv. avg.)	0.14	0.14	0.74	0.25	0.25	0.69	0.39	0.49	0.53	0.40	0.38	0.13	0.65	0.59	0.27	0.41	0.65	0.42	0.42
	E4 (mv. avg.) vs. VP (mv. avg.)	0.42	0.71	0.30	0.57	0.38	0.64	0.57	0.41	0.25	0.55	0.93	0.65	0.38	0.43	0.50	0.47	0.50	0.83	0.53
	E4 filtered vs. VP filtered	0.24	0.16	0.80	0.42	0.40	0.80	0.47	0.64	0.73	0.48	0.44	0.28	0.78	0.66	0.29	0.50	0.84	0.73	0.54
	E4 vs. VP	0.51	0.74	0.32	0.58	0.40	0.66	0.64	0.48	0.27	0.57	0.96	0.77	0.41	0.46	0.64	0.69	0.52	0.91	0.58
	Total AVG	0.33	0.44	0.54	0.46	0.36	0.70	0.52	0.50	0.45	0.50	0.68	0.46	0.55	0.54	0.42	0.51	0.63	0.72	0.52
VLF ECG	BH (mv. avg.) vs. VP (mv. avg.)	0.11	0.17	0.16	0.09	0.09	0.10	0.12	0.08	0.14	0.11	0.11	0.45	0.29	0.49	0.74	0.54	0.72	0.73	0.29
	BH vs. VP	0.11	0.17	0.31	0.10	0.29	0.16	0.16	0.10	0.21	0.35	0.13	0.59	0.43	0.77	0.95	0.73	0.96	0.91	0.41
	Total AVG	0.11	0.17	0.23	0.10	0.19	0.13	0.14	0.09	0.17	0.23	0.12	0.52	0.36	0.63	0.85	0.64	0.84	0.82	0.35
LF ECG	BH (mv. avg.) vs. VP (mv. avg.)	0.05	0.06	0.17	0.09	0.07	0.09	0.12	0.10	0.14	0.04	0.10	0.21	0.25	0.51	0.74	0.36	0.75	0.79	0.26
	BH vs. VP	0.10	0.10	0.30	0.10	0.28	0.11	0.16	0.12	0.21	0.06	0.13	0.30	0.32	0.72	1.00	0.54	0.97	0.93	0.36
	Total AVG	0.08	0.08	0.24	0.10	0.17	0.10	0.14	0.11	0.17	0.05	0.11	0.26	0.28	0.61	0.87	0.45	0.86	0.86	0.31
HF ECG	BH (mv. avg.) vs. VP (mv. avg.)	0.05	0.08	0.15	0.10	0.09	0.08	0.13	0.07	0.14	0.05	0.14	0.05	0.35	0.28	0.58	0.45	0.45	0.48	0.21
	BH vs. VP	0.20	0.13	0.44	0.12	0.16	0.11	0.22	0.25	0.23	0.17	0.50	0.17	0.76	0.84	0.91	1.00	0.99	0.88	0.45
	Total AVG	0.13	0.10	0.30	0.11	0.13	0.10	0.17	0.16	0.18	0.11	0.32	0.11	0.56	0.56	0.74	0.72	0.72	0.68	0.33

Figure 12. Global Fréchet distance matrix of pairs of parameter and participants; detail (**a**) complements Figure 4, detail (**b**) complements Figure 5; color: green indicates low distance thus high similarity, blue indicates high distance thus low similarity.

In addition to the global geometric similarity, Figure 13 shows local similarity characteristics of the time series using a moving windows approach. The figure shows that the local Fréchet distance of a 1-min moving window indeed reveals differences in similarity at different intensities of physical activity (0–300 s: no activity; 301–900 s: cycling with increasing intensity; 901–1200 s: no activity–cool down; for details refer to Section 4.1). For instance, IBI derived from ECG tends to have a rather constant similarity over the entire measurement period (Figure 13, fourth row), and it tends to be more similar than IBI measured "directly" (Figure 13, first and second row).

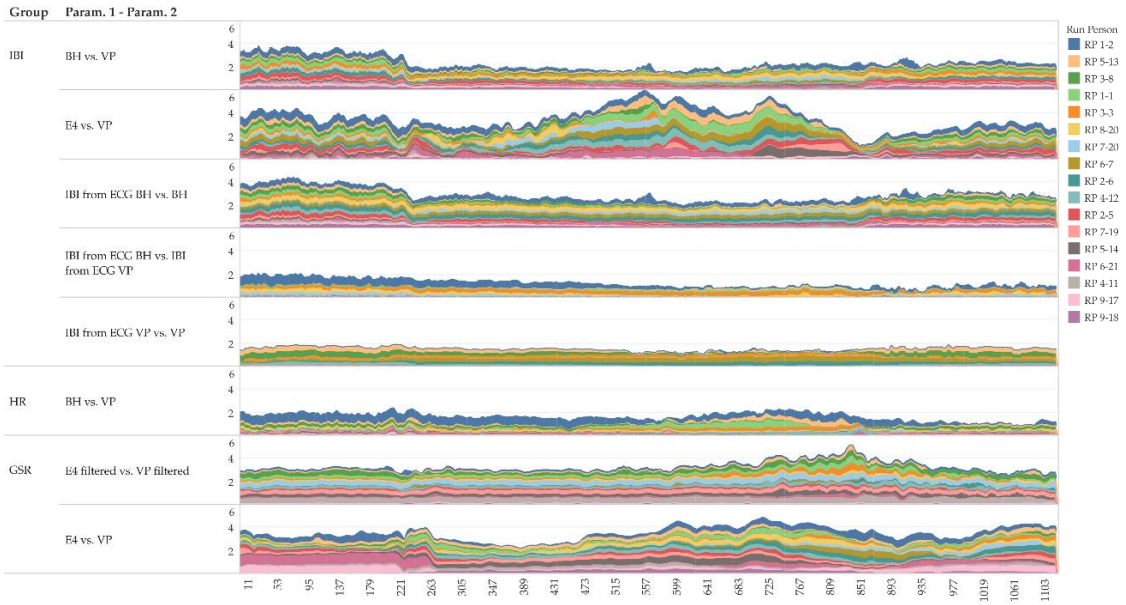

Figure 13. Local Fréchet distance of a moving time window (1 min) of selected pairs of parameters (inter beat interval IBI, heart rate HR, galvanic skin response GSR; moving averaged only).

5.2.5. DTW Distance

The Dynamic Time Warping (DTW) distance is a measure typically used to assess the similarity of time series [50,51]. Simply speaking, DTW tries to optimize the alignment of one time series (test) with another (reference) by stretching or shrinking it in a non-linear fashion along its time axis. The overall distance is the sum of all distances between pairs of points. Identical time series have a distance of zero. Figure 14 shows the DTW distance between pairs of parameters and participants, detail (a) refers to the HR example shown in Figure 4 and detail (b) refer to the GSR example shown in Figure 5.

Group	Parameter 1 - Parameter 2	RP 5-14	RP 4-11	RP 9-18	RP 4-12	RP 6-21	RP 6-7	RP 2-6	RP 2-5	RP 7-19	RP 8-19	RP 9-17	RP 7-20	RP 8-20	RP 3-3	RP 1-1	RP 5-13	RP 3-8	RP 1-2	Total AVG
IBI	IBI from ECG BH vs. IBI from ECG VP	9.9	16.4	20.5	15.1	14.0	5.1	11.7	18.9	28.3	120.7	14.2	85.1	107.8	169.3	59.9	53.5	16.8	274.5	57.9
	IBI form ECG BH (mv. avg.) vs. IBI from ECG VP (mv. avg.)	2.4	4.2	4.9	4.2	4.3	1.5	4.8	4.9	11.6	191.7	5.4	99.2	125.0	159.8	41.8	35.7	4.0	359.4	59.2
	IBI from ECG VP (mv. avg.) vs. VP (mv. avg.)	4.8	4.5	5.7	77.7	4.1	219.5	123.8	6.0	14.5	1.3	13.2	11.6	5.5	95.2	6.8	247.6	279.1	7.9	62.7
	IBI from ECG VP vs. VP	18.6	14.8	24.1	76.6	11.0	180.6	84.9	22.0	32.5	7.1	22.6	26.1	16.2	135.5	19.2	227.4	264.6	14.6	67.7
	BH vs. VP	58.5	61.8	163.3	63.6	63.1	53.5	93.9	166.1	71.4	151.0	98.7	98.0	100.6	135.7	170.5	110.3	106.9	191.2	108.9
	IBI from ECG BH vs. BH	57.4	61.6	165.6	88.9	63.3	194.7	189.0	167.2	72.9	48.6	95.8	83.8	147.3	110.5	133.0	101.0	118.0	131.4	112.8
	E4 vs. VP	113.7	39.4	60.6	98.4	102.9	141.9	99.0	106.5	67.3	99.1	64.1	106.9	62.0	76.1	156.6	133.1	158.1	356.4	113.6
	BH (mv. avg.) vs. VP (mv. avg.)	31.5	45.3	153.1	26.1	67.9	39.9	93.3	161.4	44.3	335.2	72.1	128.2	140.6	165.2	192.9	101.4	79.2	289.5	120.4
	E4 (mv. avg.) vs. VP (mv. avg.)	139.1	15.6	30.5	104.8	121.6	178.2	122.7	105.7	59.9	140.5	32.5	120.6	59.7	52.5	240.6	205.4	179.5	374.7	126.9
	IBI from ECG (mv. avg.) vs. BH (mv. avg.)	30.3	45.5	153.5	107.2	66.5	170.8	241.1	164.2	48.5	28.5	60.5	126.2	369.9	78.5	141.1	108.2	130.9	243.3	128.6
	Total AVG	46.6	30.9	78.4	66.2	51.9	118.6	106.4	92.3	45.1	112.4	47.9	88.6	113.5	119.8	116.2	132.6	133.7	224.3	95.9
HR	BH (mv. avg.) vs. VP (mv. avg.)	20.3	10.6	13.7	47.1	19.7	7.5	6.3	8.1	34.5	136.4	30.0	17.7	41.9	74.1	37.8	121.4	117.4	261.3	55.9
	BH vs. VP	31.0	22.6	32.1	55.9	26.7	20.1	15.0	27.9	51.0	104.7	37.7	34.7	61.5	74.0	47.7	128.0	116.9	288.1	65.3
	Total AVG	25.7	16.6	22.9	51.5	23.2	13.8	10.8	18.0	42.8	120.5	33.8	26.2	51.7	74.1	42.8	124.7	117.2	274.7	60.6
GSR	E4 filtered vs. VP filtered	93.5	121.1	27.4	80.4	43.9	27.9	34.0	60.7	354.6	12.5	58.5	92.1	55.5	99.5	91.8	45.3	156.8	3.2	84.0
	E4 vs. VP	28.4	18.7	39.4	69.7	294.3	26.4	92.5	144.6	101.7	60.2	409.9	51.1	75.5	62.5	56.0	74.8	39.7	260.7	105.9
	E4 (mv. avg.) vs. VP (mv. avg.)	25.6	15.2	33.8	68.8	295.5	25.9	91.6	142.0	112.0	79.1	432.9	60.8	74.5	59.9	61.8	69.7	42.8	289.2	110.1
	E4 filtered (mv. avg.) vs. VP filtered (mv. avg.)	102.9	286.1	22.9	91.7	40.4	20.2	33.5	68.7	423.3	11.4	96.6	319.4	31.6	97.6	86.4	43.8	191.9	46.2	113.2
	Total AVG	62.6	110.3	30.9	77.6	168.5	25.1	62.9	104.0	248.5	40.8	249.5	130.9	64.3	79.9	74.0	58.4	107.8	163.1	103.3
VLF ECG	BH (mv. avg.) vs. VP (mv. avg.)	8.6	8.6	18.2	5.6	7.7	3.9	8.4	14.8	5.4	19.1	23.2	30.5	70.0	23.9	42.8	40.7	16.3	83.4	24.0
	BH vs. VP	12.6	17.1	31.5	10.4	12.0	7.0	15.5	27.5	11.7	27.0	35.2	43.0	71.0	33.0	58.5	56.7	27.9	87.2	32.5
	Total AVG	10.6	12.8	24.9	8.0	9.8	5.4	11.9	21.1	8.5	23.1	29.2	36.8	70.5	28.5	50.7	48.7	22.1	85.3	28.2
LF ECG	BH (mv. avg.) vs. VP (mv. avg.)	3.6	8.9	7.8	6.6	6.4	3.3	6.9	13.1	7.5	9.0	8.6	49.1	91.7	46.4	73.0	32.2	8.9	99.0	26.9
	BH vs. VP	7.6	15.2	15.3	18.7	12.9	7.0	12.7	20.6	13.6	14.6	18.2	61.6	86.3	49.5	73.2	41.8	21.5	100.5	32.8
	Total AVG	5.6	12.0	11.6	12.7	9.6	5.2	9.8	16.8	10.5	11.8	13.4	55.3	89.0	47.9	74.1	37.0	13.2	99.8	29.9
HF ECG	BH (mv. avg.) vs. VP (mv. avg.)	7.4	7.1	9.1	8.6	10.1	6.8	8.2	8.2	4.1	4.6	4.6	40.2	41.5	41.9	101.7	33.9	10.2	84.7	24.1
	BH vs. VP	19.4	27.9	39.7	39.5	32.0	23.0	41.4	29.0	14.6	18.8	18.1	46.3	55.0	55.9	85.5	56.0	38.1	97.8	41.0
	Total AVG	13.4	17.5	24.4	24.0	21.0	14.9	24.8	18.6	9.4	11.7	11.3	43.2	48.2	48.9	93.6	45.0	24.1	91.3	32.5

Figure 14. Dynamic Time Warping (DTW) distance matrix of pairs of parameter and participants; detail (**a**) complements Figure 4, detail (**b**) complements Figure 5; colour: green indicates low distance thus high similarity, purple indicates high distance thus low similarity.

The global DTW distances from Figure 14 can be illustrated as an individual pairwise comparison of time series. For instance, Figure 15 shows an example of the DTW distance between two time series of one participant's HR measurements, which are highly similar (low DTW distance). The corresponding exploratory plots of the HR example are shown in Figure 4**.** Figure 16 shows an example of two time

series of GSR measurements with rather low similarity (high DTW distance); however, the overall trend is highly similar. The corresponding exploratory plots of the GSR example are shown in Figure 5.

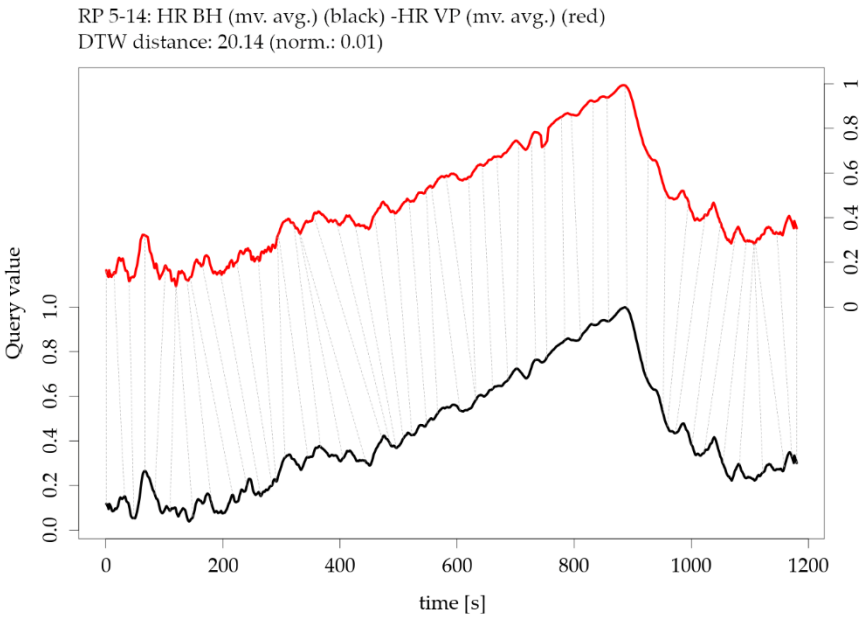

Figure 15. Illustration of the Dynamic Time Warping (DTW) distance between two parameters of participant RP 5-14: moving averaged version of heart rate HR from BioHarness BH versus moving averaged version of heart rate HR from VarioPort VP (note the offset of the two y-axes).

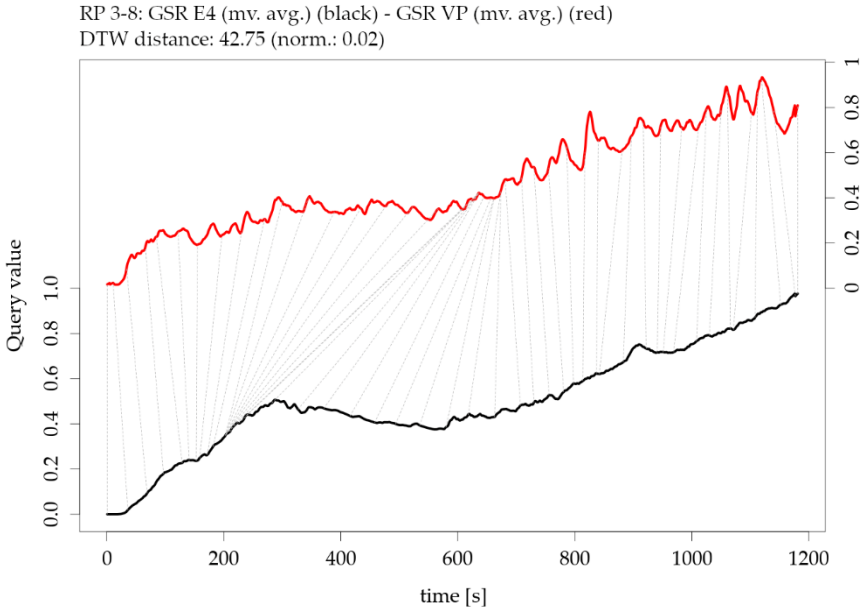

Figure 16. Illustration of the Dynamic Time Warping (DTW) distance between two parameters of participant RP 3-8: moving averaged version of galvanic skin response GSR from E4 versus moving averaged version of galvanic skin response GSR from VarioPort VP (note the offset of the two y-axes).

6. Discussion and Limitations

Overall, the sensor benchmarking worked well, both from a standardized laboratory study and data acquisition viewpoint, as well as from the data analysis methodology perspective. The high correlations between the cardiovascular parameters HR and IBI were as expected because these parameters are comparably simple to measure through a range of methodologies (electrical, optical).

The high correlations between the other ECG-derived measurements were a little more surprising because (1) ECG is measured through a multi-channel electric current-based system, which is a complex procedure; (2) the use of contact electrodes of the BioHarness sensor (in contrast to the sticky electrodes of high-quality sensors) may cause contact (and thus measurement-) problems; 3) ECG is measured at a very high frequency (at least 200 Hz), which is technologically challenging for low-cost wearables.

For GSR, our experiment resulted in lower, but still reasonable similarities, which may be caused by a number of factors like different measurement methods (sticky electrodes vs. plate electrode), and different placement of the sensors (hand palm vs. wrist), etc.

From a more general point of view, it is a known issue that low-cost wearable sensors tend to be prone to producing datasets that suffer from reduced data quality—even though we checked the appropriate positioning of the sensors before we started the exercise.

A particular issue arose with participant RP 1–2. As the results show, the measurements for this participant indicate low correlations for almost all physiological parameters. This may be due to problems with the contact between the electrodes and the skin, which may have been compromised by the person's physical characteristics.

A vital part of the analysis is the visualisation of results on two complementary levels: First, on the individual level, the data from different sensors measuring the same physiological parameter at the same time on the same participant provide a direct comparison between the two-time series of interest. This enables reaching conclusions on the sensors' measuring behaviour. Second, on the collective level, the consolidation of global metrics allows for comparing signals between participants. This further provides useful insights into the influence of the participants' individual components (physical constitution, individual baseline level of skin conductance, etc.).

These complementary visualizations enable a flexible method of interpretation. For instance, it allows starting the interpretation on the individual level on a particular pair of physiological parameters of interest (e.g., HR of participant RP 5–14) using the corresponding exploratory plot as shown in Figure 4, then rolling-up using the R^2 matrix (Figure 6a) together with the cross-correlation matrix (Figure 7a) and comparing the individual result between participants. Further, it allows checking whether that particular pair of parameters has a functional relationship and whether that relationship is stable among other participants by using the MIC–R^2 individual matrix (Figure 11). The focus on a particular pair of parameters and participant can be continued to the similarity measures, namely the Fréchet distance and the DTW distance (Figures 12 and 14). Another method of interpretation is to begin at the collective level using the MIC–R^2 cluster matrix of pairs of parameters (Figure 10), then drilling-down on a specific parameter combination of interest using the MIC–R^2 individual matrix (Figure 11) and contextualizing this matrix with the corresponding exploratory plots as shown in Figure 5.

This kind of visualisation provides the central advantage regarding the sensor benchmarking from a "big picture" view, i.e., to serve as a basis for visual analysis of the correlations between the measurements of one parameter as measured by two different sensors (each row in the matrix) and the correlations between the different parameters for a single participant (each column in the matrix). Furthermore, the matrix allows the simple assessment of each single cell to trace back particularities of each measurement to a test person, which makes it easier to single out anomalies that may be caused by usage errors, a user's characteristics, single sensor failures, or violations of the benchmark protocol.

The cross-correlation analysis shows that groups of physiological parameters can be associated with different patterns of temporal shifts. As illustrated in Figure 7, the cross-correlation also varies between participants: from overall positive for IBI derived from ECG for more than 60% of the participants to highly positive at small lags and highly negative at larger lags at the heart rate variability parameters VLF, LF, and HF. Although the clocks of the sensors were synchronized right before the study, we observed a lag of 1–2 s between HR measurements (BH versus VP) and GSR measurements (E4 versus VP), as exemplarily shown in Figure 4a,c and Figure 5c,d, respectively. In all

cases, the VP time series were "leading", which may indicate that the response characteristics of the VP sensors are more sensitive compared to the other sensors.

Measuring the strength of association between pairs of physiological parameters was of particular interest. Herein, we contrasted the coefficient of determination R^2 with the MIC (Figures 10 and 11). In other words, we confronted a statistic that measures linear relationships against a statistic that measures all types of functional relationships, including linear ones, and thereby classified the relationship as 'false linear', 'true linear', or 'functional but not linear'. The results are outstanding: on the one hand, some already expected linear relationships have been confirmed by a purely data-driven approach (for instance, relationships between IBI and VLF, LF, HF); on the other hand, some relationships that were expected to be linear are in fact not linear or functional. For instance, the relationships between GSR measured by E4 and GSR measured by VP (both filtered and moving averaged versions).

7. Conclusions and Future Work

In this paper, we performed a benchmark of two wearable physiological sensors (Zephyr BioHarness 3 and Empatica E4) by comparing their measurements (heart rate, inter-beat interval, and galvanic skin response, and derived heart rate variability parameters) to highly-calibrated high-end professional equipment. In our study, we used the measurements from 18 participants to compare the correlations (Pearson's r), cross-correlations at different temporal lags from −15 sec to +15 sec, the (sub-)linearity of functional dependencies (MIC), the difference of two measurement time series with respect to their geometric structure (Fréchet distance), local time series similarities (moving window), and time series similarity with respect to their temporal alignment (DTW).

The results of our study show that the measured cardiovascular parameters yield very high similarities between the low-cost wearable and the calibrated professional sensors. Although cardiovascular parameters are simple to measure (technically and phenomenon-wise), the obtained similarities are remarkable. For GSR, our experiment resulted in lower similarities, which may be caused by a number of factors like different measurement methods, different placement of the sensors (hand palm vs. wrist), conduction characteristics between skin and sensor surface (use of electrolyte gel or not), and others. It should be noted that the use of isotonic electrolyte gel is a scientific standard for measurement of electrodermal activity [52] and was used with the Varioport GSR measure but not with the other devices.

We demonstrated that our methodological approach to quantify correlations and similarities on both the individual and the aggregated level can provide interesting insights into the relationships between and among physiological parameters. The many figures generated (only the most essential ones are presented in this paper) enable different points of view on the same data and thus a more holistic interpretation for the benchmark of physiological sensors. Our research contributes to such a holistic interpretation in two ways: 1) the confrontation of the coefficient of determination R^2 against the Maximal Information Coefficient MIC, in particular, the classification of non-linear correlations, and 2) the quantification of the signals' temporal and geometric similarity based on well-established distance metrics (DTW distance and Fréchet distance).

Our future work will focus on two main research challenges. First, to continue fine-tuning the methodology and integrate additional similarity measures, for instance, the Time Warp Edit Distance (TWED) [53]. Second, to evaluate the transferability of the methodology to other time series benchmarking challenges, not necessarily physiological measurements. In the long run we want to expand the methodology to the geospatial domain, i.e., integrating the location in addition to the timestamp and the measurement of mobile sensors. This approach will likely warrant an additional field study that addresses the suitability of measurement devices and measurement quality on moving subjects, e.g., persons riding a bicycle or walking, and relating sensor data to subjective experience self-report data.

Author Contributions: Conceptualization, G.S., B.R., A.P., K.K., M.L. and F.H.W.; Data curation, G.S., B.R., A.P., K.K. and M.L.; Formal analysis, G.S. and B.R.; Investigation, G.S., B.R., A.P. and M.L.; Methodology, G.S., B.R.

and K.K.; Project administration, A.P.; Software, G.S., A.P., K.K. and M.L.; Validation, G.S., B.R., M.L. and F.H.W.; Visualisation, G.S.; Writing—Original draft, G.S., B.R. and M.L.; Writing—Review & editing, G.S., B.R., A.P., K.K., M.L. and F.H.W.

Funding: This research was supported by the Austrian Science Fund (FWF) through the project "Urban Emotions" (FWF I-3022) and by the Austria Research Promotion Agency (FFG) through the project "Walk&Feel" (FFG 865208). This research was partly funded by the Austrian Science Fund (FWF) through the Doctoral College GIScience (DK W 1237-N23).

Acknowledgments: Open Access Funding by the Austrian Science Fund (FWF). We would like to thank all participants of the case study who made this work possible by donating several hours of their free time.

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

References

1. Swan, M. Sensor mania! The internet of things, wearable computing, objective metrics, and the quantified self 2.0. *J. Sens. Actuator Netw.* **2012**, *1*, 217–253. [CrossRef]
2. Crawford, K.; Lingel, J.; Karppi, T. Our metrics, ourselves: A hundred years of self-tracking from the weight scale to the wrist wearable device. *Eur. J. Cult. Stud.* **2015**, *18*, 479–496. [CrossRef]
3. Piwek, L.; Ellis, D.A.; Andrews, S.; Joinson, A. The rise of consumer health wearables: Promises and barriers. *PLoS Med.* **2016**, *13*, e1001953. [CrossRef]
4. Werner, C.; Resch, B.; Loidl, M. Evaluating urban bicycle infrastructures through intersubjectivity of stress sensations derived from physiological measurements. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 265. [CrossRef]
5. Basu, S.; Jana, N.; Bag, A.; Mahadevappa, M.; Mukherjee, J.; Kumar, S.; Guha, R. Emotion recognition based on physiological signals using valence-arousal model. In Proceedings of the 2015 Third International Conference on Image Information Processing (ICIIP), Waknaghat, India, 21–24 December 2015; pp. 50–55.
6. Resch, B.; Summa, A.; Sagl, G.; Zeile, P.; Exner, J.-P. Urban emotions—Geo-semantic emotion extraction from technical sensors, human sensors and crowdsourced data. In *Progress in Location-Based Services 2014*; Gartner, G., Huang, H., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 199–212.
7. Taj-Eldin, M.; Ryan, C.; O'Flynn, B.; Galvin, P. A review of wearable solutions for physiological and emotional monitoring for use by people with autism spectrum disorder and their caregivers. *Sensors* **2018**, *18*, 4271. [CrossRef]
8. Healey, J. Physiological sensing of emotion. In *The Oxford Handbook of Affective Computing*; Oxford University Press: Oxford, UK, 2014; p. 204.
9. Peake, J.M.; Kerr, G.; Sullivan, J.P. A critical review of consumer wearables, mobile applications, and equipment for providing biofeedback, monitoring stress, and sleep in physically active populations. *Front. Physiol.* **2018**, *9*, 743. [CrossRef] [PubMed]
10. Birenboim, A.; Dijst, M.; Scheepers, F.E.; Poelman, M.P.; Helbich, M. Wearables and location tracking technologies for mental-state sensing in outdoor environments. *Prof. Geogr.* **2019**, *71*, 449–461. [CrossRef]
11. Kyriakou, K.; Resch, B.; Sagl, G.; Petutschnig, A.; Werner, C.; Niederseer, D.; Liedlgruber, M.; Wilhelm, F.H.; Osborne, T.; Pykett, J. Detecting moments of stress from measurements of wearable physiological sensors. *Sensors* **2019**, *19*, 3805. [CrossRef] [PubMed]
12. Zeile, P.; Resch, B. Combining biosensing technology and virtual environments for improved urban planning. *GI_Forum* **2018**, *1*, 344–357. [CrossRef]
13. Dörrzapf, L.; Kovács-Győri, A.; Resch, B.; Zeile, P. Defining and assessing walkability: A concept for an integrated approach using surveys, biosensors and geospatial analysis. *Urban Dev. Issues* **2019**, *62*, 5–15. [CrossRef]
14. Guo, R.; Li, S.; He, L.; Gao, W.; Qi, H.; Owens, G. Pervasive and unobtrusive emotion sensing for human mental health. In Proceedings of the 2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops, Venice, Italy, 5–8 May 2013; pp. 436–439.
15. Majumder, S.; Mondal, T.; Deen, M. Wearable sensors for remote health monitoring. *Sensors* **2017**, *17*, 130. [CrossRef] [PubMed]
16. Kenry, Y.J.C.; Lim, C.T. Emerging flexible and wearable physical sensing platforms for healthcare and biomedical applications. *Microsyst. Nanoeng.* **2016**, *2*, 16043. [CrossRef] [PubMed]
17. Giakoumis, D.; Tzovaras, D.; Hassapis, G. Subject-dependent biosignal features for increased accuracy in psychological stress detection. *Int. J. Hum. Comput. Stud.* **2013**, *71*, 425–439. [CrossRef]

18. Gradl, S.; Wirth, M.; Richer, R.; Rohleder, N.; Eskofier, B.M. An overview of the feasibility of permanent, real-time, unobtrusive stress measurement with current wearables. In Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare, ACM, Trento, Italy, 20–23 May 2019; pp. 360–365.
19. Aghabozorgi, S.; Shirkhorshidi, A.S.; Wah, T.Y. Time-series clustering—A decade review. *Inf. Syst.* **2015**, *53*, 16–38. [CrossRef]
20. Serrà, J.; Arcos, J.L. An empirical evaluation of similarity measures for time series classification. *Knowl. Based Syst.* **2014**, *67*, 305–314. [CrossRef]
21. Shin, K. An alternative approach to measure similarity between two deterministic transient signals. *J. Sound Vib.* **2016**, *371*, 434–445. [CrossRef]
22. Toohey, K.; Duckham, M. Trajectory similarity measures. *Sigspatial Spec.* **2015**, *7*, 43–50. [CrossRef]
23. Wang, X.; Mueen, A.; Ding, H.; Trajcevski, G.; Scheuermann, P.; Keogh, E. Experimental comparison of representation methods and distance measures for time series data. *Data Min. Knowl. Discov.* **2013**, *26*, 275–309. [CrossRef]
24. Chen, L.; Özsu, M.T.; Oria, V. Robust and fast similarity search for moving object trajectories. In Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, Baltimore, MD, USA, 14–16 June 2005; pp. 491–502.
25. Keogh, E.; Ratanamahatana, C.A. Exact indexing of dynamic time warping. *Knowl. Inf. Syst.* **2005**, *7*, 358–386. [CrossRef]
26. Folgado, D.; Barandas, M.; Matias, R.; Martins, R.; Carvalho, M.; Gamboa, H. Time alignment measurement for time series. *Pattern Recognit.* **2018**, *81*, 268–279. [CrossRef]
27. Jiang, G.; Wang, W.; Zhang, W. A novel distance measure for time series: Maximum shifting correlation distance. *Pattern Recognit. Lett.* **2019**, *117*, 58–65. [CrossRef]
28. Kate, R.J. Using dynamic time warping distances as features for improved time series classification. *Data Min. Knowl. Discov.* **2016**, *30*, 283–312. [CrossRef]
29. Fréchet, M.M. Sur quelques points du calcul fonctionnel. *Rendiconti del Circolo Matematico di Palermo (1884–1940)* **1906**, *22*, 1–72. [CrossRef]
30. Shahbaz, K. Applied similarity problems using fréchet distance. *arXiv* **2013**, arXiv:preprint/1307.6628.
31. De Carufel, J.-L.; Gheibi, A.; Maheshwari, A.; Sack, J.-R.; Scheffer, C. Similarity of polygonal curves in the presence of outliers. *Comput. Geom.* **2014**, *47*, 625–641. [CrossRef]
32. Aronov, B.; Har-Peled, S.; Knauer, C.; Wang, Y.; Wenk, C. Fréchet distance for curves, revisited. In *European Symposium on Algorithms*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 52–63.
33. Wylie, T.R. The Discrete Fréchet Distance with Applications. Ph.D. Thesis, Montana State University-Bozeman, College of Engineering, Bozeman, MT, USA, 2013.
34. Kianimajd, A.; Ruano, M.G.; Carvalho, P.; Henriques, J.; Rocha, T.; Paredes, S.; Ruano, A.E. Comparison of different methods of measuring similarity in physiologic time series. *IFAC-PapersOnLine* **2017**, *50*, 11005–11010. [CrossRef]
35. Hauke, J.; Kossowski, T. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaest. Geogr.* **2011**, *30*, 87–93. [CrossRef]
36. Reshef, D.N.; Reshef, Y.A.; Finucane, H.K.; Grossman, S.R.; McVean, G.; Turnbaugh, P.J.; Lander, E.S.; Mitzenmacher, M.; Sabeti, P.C. Detecting novel associations in large data sets. *Science* **2011**, *334*, 1518–1524. [CrossRef]
37. Speed, T. A correlation for the 21st century. *Science* **2011**, *334*, 1502. [CrossRef] [PubMed]
38. Morelli, M.S.; Greco, A.; Valenza, G.; Giannoni, A.; Emdin, M.; Scilingo, E.P.; Vanello, N. Analysis of generic coupling between EEG activity and $P_{ET}CO_2$ in free breathing and breath-hold tasks using maximal information coefficient (MIC). *Sci. Rep.* **2018**, *8*, 4492. [CrossRef]
39. Brillinger, D.R. *Time Series: Data Analysis and Theory*; Siam: San Francisco, CA, USA, 2001; Volume 36.
40. Dickey, D.A.; Fuller, W.A. Distribution of the estimators for autoregressive time series with a unit root. *J. Am. Stat. Assoc.* **1979**, *74*, 427–431.
41. Johnstone, J.A.; Ford, P.A.; Hughes, G.; Watson, T.; Garrett, A.T. Bioharness (™) multivariable monitoring device: Part. II: Reliability. *J. Sports Sci. Med.* **2012**, *11*, 409–417. [PubMed]
42. Johnstone, J.A.; Ford, P.A.; Hughes, G.; Watson, T.; Garrett, A.T. Bioharness (™) multivariable monitoring device: Part. I: Validity. *J. Sports Sci. Med.* **2012**, *11*, 400–408. [PubMed]

43. Blechert, J.; Peyk, P.; Liedlgruber, M.; Wilhelm, F.H. Anslab: Integrated multichannel peripheral biosignal processing in psychophysiological science. *Behav. Res. Methods* **2016**, *48*, 1528–1545. [CrossRef] [PubMed]
44. Bluemke, M.; Resch, B.; Lechner, C.; Westerholt, R.; Kolb, J.-P. Integrating geographic information into survey research: Current applications, challenges and future avenues. *Surv. Res. Methods* **2017**, *11*, 307–327.
45. Bar-Joseph, Z.; Gerber, G.K.; Gifford, D.K.; Jaakkola, T.S.; Simon, I. Continuous representations of time-series gene expression data. *J. Comput. Biol.* **2003**, *10*, 341–356. [CrossRef]
46. Wilhelm, F.H.; Grossman, P.; Roth, W.T. Assessment of heart rate variability during alterations in stress: Complex demodulation vs. Spectral analysis. *Biomed. Sci. Instrum.* **2005**, *41*, 346–351.
47. Li, L.; Caldwell, G.E. Coefficient of cross correlation and the time domain correspondence. *J. Electromyogr. Kinesiol.* **1999**, *9*, 385–389. [CrossRef]
48. Reshef, D.N.; Reshef, Y.A.; Sabeti, P.C.; Mitzenmacher, M. An empirical study of the maximal and total information coefficients and leading measures of dependence. *Ann. Appl. Stat.* **2018**, *12*, 123–155. [CrossRef]
49. Alt, H. The computational geometry of comparing shapes. In *Efficient Algorithms: Essays Dedicated to Kurt Mehlhorn on the Occasion of His 60th Birthday*; Albers, S., Alt, H., Näher, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2009; pp. 235–248.
50. Zhu, Q.; Batista, G.; Rakthanmanon, T.; Keogh, E. A novel approximation to dynamic time warping allows anytime clustering of massive time series datasets. In Proceedings of the 2012 SIAM International Conference on Data Mining, Davis, FL, USA, 26–28 April 2012; pp. 999–1010.
51. Tormene, P.; Giorgino, T.; Quaglini, S.; Stefanelli, M. Matching incomplete time series with dynamic time warping: An algorithm and an application to post-stroke rehabilitation. *Artif. Intell. Med.* **2009**, *45*, 11–34. [CrossRef]
52. Fowles, D.C.; Christie, M.J.; Edelberg, R.; Grings, W.W.; Lykken, D.T.; Venables, P.H. Publication recommendations for electrodermal measurements. *Psychophysiology* **1981**, *18*, 232–239. [CrossRef] [PubMed]
53. Marteau, P.-F. Time warp edit distance with stiffness adjustment for time series matching. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 306–318. [CrossRef] [PubMed]