Journal of Computational Science 88 (2025) 102602

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

Generating vertical ground reaction forces using a stochastic data-driven model for pedestrian walking

Alvaro Magdaleno [a], José María García-Terán [a], César Peláez-Rodríguez [a,b], Guillermo Fernández [a], Antolin Lorenzana [a,*]

[a] ITAP. Universidad de Valladolid, Paseo del Cauce 59, Valladolid, 47011, Castilla y Leon, Spain
[b] Universidad de Alcalá, Departamento de Teoría de la Señal y Comunicaciones, Crta. Madrid-Barcelona km. 33, 600, Alcalá de Henares, 28805, Madrid, Spain

ARTICLE INFO

Keywords:
Human loading
Walking load model
Stochastic data-driven model
Virtual GRFs

ABSTRACT

A novel time-domain approach to the characterization of the forces induced by a pedestrian is proposed. It focuses on the vertical component while walking, but thanks to how it is conceived, the algorithm can be easily adapted to other activities or any other force component. The work has been developed from the statistical point of view, so a stochastic data-driven model is finally obtained after the algorithm is applied to a set of experimentally measured steps. The model is composed of two mean vectors and their corresponding covariance matrices to represent the steps, as well as some more means and standard deviations to account for the step scaling and double support phase, under the assumption that the random variables follow normal distributions. Velocity and step length are also provided, so the model and the latter data enable the realistic generation of virtual gaits. Some application examples at different walking paces are shown, in which comparisons between the original steps and a set of virtual ones are performed to show the similarities between both. For reproducibility purposes, the data and the developed algorithm have been made available.

1. Introduction

Walking is quite a complex process that involves numerous human abilities, such as balance and force, required to make a person advance while overcoming gravity, terrain irregularities, and air friction. This process is activated not only for traversing straight paths, but also for making closed turns, avoiding obstacles, and going up and down stairs or slopes. The movement is produced by the force that people develop with their feet and legs. The forces exerted on the ground to facilitate movement are typically called Ground Reaction Forces (GRFs) and have three components [1]: a longitudinal one, which propels the person forward; a transversal one, associated with the natural sway movement occurring during the transition from one foot to another; and a vertical one, linked to the pedestrian's weight and the feet leaving and returning to the ground. The complexity of walking makes it a non-deterministic action, meaning the forces developed in a particular step under certain circumstances are statistically similar but not identical to those of the previous or next step, making accurate modeling and prediction challenging due to the necessity of not neglecting the natural randomness of gait. Characterizing this action is crucial in areas such as biomechanics, physiotherapy, and structural engineering. Biomechanics examines limb interaction for movement like walking, aiding applications such as artificial motion replication [2] and risk assessment related to specific movements [3]. Physiotherapy focuses on walking disorders, diagnosed through force measurements over time [4]. Structural engineering, in turn, analyzes human-induced forces on the ground, with particular interest in the dominant vertical component of the GRF and its inherent variability, which this study focuses on in particular.

Many authors have dealt with the characterization of walking forces in several ways. Racic et al. [1] presented a comprehensive state of the art on this topic and divided the main techniques used to model the GRFs into two categories: time-domain and frequency-domain techniques. Time-domain techniques use the registered time series directly to estimate model parameters, typically mathematical expressions based on periodic functions, with amplitudes, frequencies, and phases optimized to represent the measured data. The model can also include velocity or step length. There are two approaches in the time domain: deterministic, where parameters are fixed values, and stochastic, where parameters are described by probability density functions to account for variability. On the other hand, frequency-domain techniques compute the frequency content and energy distribution across a frequency band of interest. The algorithm developed in this work is framed within time-domain techniques. After [1], more works

* Corresponding author.
E-mail address: antolin.lorenzana@uva.es (A. Lorenzana).

https://doi.org/10.1016/j.jocs.2025.102602
Received 6 November 2024; Received in revised form 26 March 2025; Accepted 17 April 2025
Available online 30 April 2025

focused on developing stochastic methodologies and algorithms to model the GRFs accurately that account for their randomness have been proposed. Racic et al. presented several works where vertical GRFs were modeled in various situations: in [5] for walking, in [6] for running, and in [7] for bouncing. These all used Auto Spectral Density techniques, applying Gaussian functions to preserve the random nature of the forces, avoiding the deterministic Fourier decomposition method that assumes a perfectly periodic signal. These approaches have resulted in robust stochastic models capable of reproducing the described actions, as well as step-by-step variability of a single individual, although they do not explicitily account for the potential difference between the forces of each foot analyzing each time series individually, which may have an influence in the reproduction of the resultant action.

In the same way, other authors such as Pancaldi et al. [8], presented a time-domain stochastic model for GRFs and a simplified version for comparison. The study uses both Fourier series and random variables to approximate a step: the first models step shape accurately, while the second assumes a constant force during each step and zero between steps. Continuous forces are obtained by summing individual steps. Peters et al. [9] recovered the Fourier decomposition of the vertical force and, by means of a Bayesian and frequentist approach and using a significantly large experimental dataset, proposed a stochastic model of these forces. The previous works were developed assuming that pedestrians walk on a rigid surface and, as the previous ones already cited, do not explicitily account for potential forces variability across feet. The work of Ahmadi et al. [10] proposed a novel set up, combining the GRFs measurement with the structural response, in order to show how the structural vibration affects the way people walk, as well as demonstrating the existence of a significant human–structure interaction, something that should not be disregarded when human loading is to be applied on a slender structure with low natural frequencies.

Recently, a time-domain approach based on physical models, rather than purely mathematical ones, has also been developed. Cacho and Lorenzana [11] and Lin et al. [12] presented a bi-pedal mechanical model with mass, springs, dampers, and dynamic forces, adjusted to replicate in some way human-like GRFs. The first was theoretical, while the second used registered data for fitting. Kumar et al. [13] proposed a single-degree-of-freedom oscillator model to replicate human motion. Similarly, Wang et al. [14] developed a mass–spring-damper model fitted to data from instrumented insoles for walking and bouncing actions. Koshio et al. [15] applied Inertial Measurement Units and a 3D forward dynamics model in sports activities. These studies demonstrated that simple mechanical systems can effectively reproduce pedestrian GRFs and their properties. The need for modeling pedestrians as a walking mechanism stems from the concept of human–structure interaction, where understanding the GRFs is crucial for analyzing the effects on structures. Nevertheless, although these works are interesting and robust, they mostly treat GRFs as deterministic and symmetric actions, which has already been identified as a limiting factor for a rigorous modeling of human gait.

Finally, optimization techniques and Machine Learning models have also been explored to enhance the modeling and prediction of GRFs in human walking. These methods have been applied to datasets in order to capture patterns in gait data, achieving high accuracy in GRF prediction under varying conditions. These models can complement physical models, offering improved adaptability to different scenarios [16,17]. These algorithms have shown good capabilities in generating virtual gaits that match experimental and reference data.

Considering all the topics discussed previously, this paper proposes a time-domain method for characterizing the forces exerted by a pedestrian while walking, focusing on the vertical component but adaptable to others. A stochastic, data-driven model is developed from experimentally measured steps without using Fourier decomposition or Gaussian approximations, and preserving the walking action's randomness. Each foot's GRF is processed individually, with the resultant total action evaluated at the end, taking into account explicit differences across feet. The method detects and characterizes steps from raw force data, deriving mean vectors and covariance matrices for each foot, assuming a multivariate normal distribution. This enables the generation of random steps through a purely data-driven stochastic procedure. Mean values and standard deviations represent step scaling and the double support phase as univariate normal distributions, enabling virtual GRF generation. Examples of applications at different walking paces are provided, comparing the original data with the virtual steps.

The document is organized as follows: this Section 1 is devoted to introducing the topic and providing a thorough state of the art; Section 2 briefly presents the human gait and sets up some terminology used throughout the article; Section 3 presents the methodology, including how the experimental data were gathered and the algorithm required to process all the data; Section 4 describes the results of the application examples in terms of the characteristics of the statistical model and the virtual forces for the walking paces analyzed; finally, Section 5 summarizes the main conclusions. The end of the document is devoted to the acknowledgments and references.

2. Brief description of human gait and terminology

Before addressing the methodology, the main concepts and terminology associated to human gait are briefly presented in this section in order to ease the necessary descriptions in the following ones. As presented in the introduction, walking is a highly complex process that involves several human abilities. As a consequence, although trying to walk similarly, every individual gait is different. Nonetheless, some general characteristics are shared among a wide range of pedestrians when walking, and they are described here.

Walking is a four-phase process that periodically repeats in the so-called gait cycle. Fig. 1 shows a schema of those four phases and an example of the vertical forces that are being induced on the floor. During two phases, the double support or double limb phases (first and third in the schema), both feet are in contact with the floor. The first one starts with the left heel striking the floor and ends with the right toe off. The third stage happens in the exactly opposite way, with the right heel striking the floor and ends with the left toe off. In between, two swing or aerial phases occur: during the second phase the right foot is not in contact with the floor; while during the fourth phase, it is the left foot which is swaying in the air.

The typical vertical GRFs applied during a gait cycle are shown under the walking schema in Fig. 1. As can be seen, in both left and right feet forces, a local maximum (P_1) is found right after a heel strikes the floor, a process during which a significant fraction of the body weight is loaded on that foot for a very short time. Another local maximum (P_3) occurs during a toe off due to the impulse made by that foot to keep the forward movement. A local minimum (P_2) is found between both maxima. Regarding the resultant force, which is the effective force applied on the ground by the pedestrian, it is important to note that the greatest maxima (P_4) occur when both feet are in contact with the floor. The height of the peak is closely related to the double support time, i.e., the time both feet forces overlap: t_{RL} when the left foot force starts before the right foot force ends, or t_{LR} in the opposite scenario. This occurs twice per gait cycle and is responsible for the primary frequency component of this force, which is the walking pace (steps/min), or walking frequency (Hz). Note that this description is valid for a comfortable walking of healthy pedestrians [18]. Other walking conditions, actions (such as running or bouncing), or pedestrian diseases may alter this schema, which the authors will explore in future works.

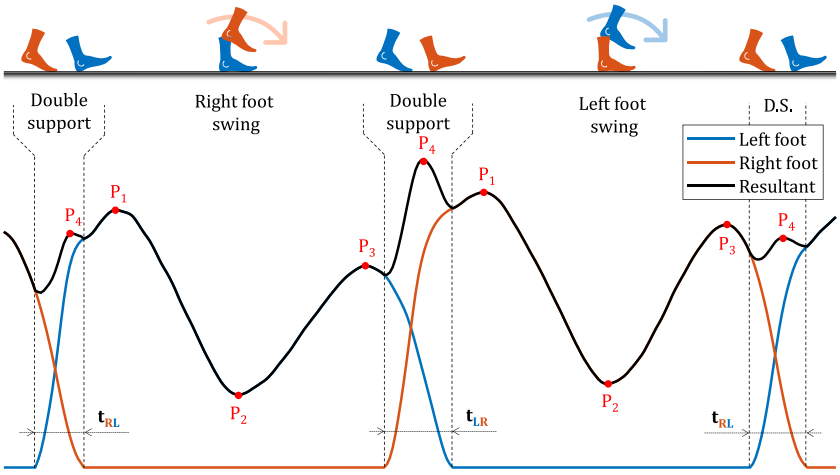


Fig. 1. Example of left (blue) and right (red) vertical GRFs together with the sum of both (black), the four phases of the gait cycle and the overlapping time variables t_{RL} and t_{LR}.

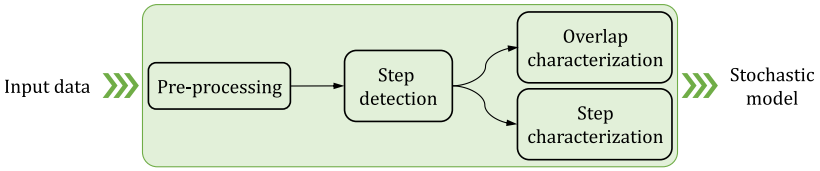


Fig. 2. Overview of the developed algorithm.

3. Methodology

This section is devoted to presenting the proposed methodology and how the experimental data for the application examples are registered. As stated in the introduction, the main objective of this work is to develop an algorithm aimed at processing the experimental data gathered while a pedestrian is walking in order to model it, so similar virtual forces can be generated.

Fig. 2 shows an overview of the developed algorithm. As can be seen, it is divided into a total of four stages: a pre-processing stage to clean the data; an automatic step detection process, to avoid an analyst to mark when every step starts and ends; the step characterization, to obtain the stochastic model of the left and right feet separately; and an overlap characterization, focused on obtaining two statistical models of the double support phase times, t_{LR} and t_{RL} separately. Each stage is detailed in subsequent sections.

3.1. GRF measurement

In this work, the vertical GRFs are measured by means of a pair of instrumented insoles. Several solutions have already been used by other authors, such as force plates [3,8,19], which do not allow forces to be continuously measured while walking along a path, or instrumented treadmills [6,20,21], which are quite bulky, expensive and complicated to calibrate and set up. This equipment is regarded as the gold standard in accuracy and test-retest reliability for GRF measurement. However, in addition to the previously aforementioned limitations, it is also restricted to controlled laboratory settings, which may influence and cause an pedestrian's gait to differ from the natural walking pattern in uncontrolled environments.

The selection of the instrumented insoles addresses the necessity of measuring GRFs over a complete and relatively long path, in a situation as close as possible to a natural, uncontrolled setting, and with acceptable accuracy compared to force plates and treadmills. Initially, custom insole-like devices were developed [22], but modern commercial solutions, such as the Loadsol® insoles from Novel GmbH (Fig. 3(a)), now offer an acceptable performance at reasonable prices. Provided in pairs, these insoles measure the resultant vertical force exerted by the foot and transmit data via Bluetooth to a smartphone datalogger at a sampling rate of 100 S/s. In this work, they are worn inside the footwear, with the battery and communication unit conveniently attached using the built-in clamping device (Fig. 3(b)).

The Loadsol® insoles have been validated for measuring vertical GRFs during activities such as walking and running, demonstrating a high correlation with force plate measurements, as well as good test-retest reliability. Additionally, studies comparing these insoles with the aforementioned force plates report mean biases of 0.6% to 3.4% and Intraclass Correlation Coefficients up to 0.97 [23,24], demonstrating an acceptable precision, though not superior to gold standard equipment. However, their ability to measure forces in real environments like flexible structures, in a situation as natural as possible, can compensate for potential deficiencies.

The registered data are used in this work to show how the algorithm is applied and the quality of the model that can be expected. In the next subsections, the algorithm is explained and application examples with real data are provided in Section 4. For reproducibility purposes, both the data and the algorithm are available online at [25].

3.2. Pre-processing

Both vectors representing the forces exerted by each foot are processed separately. Only one pre-processing operation is considered in this work, which is a selection of the relevant data; although more could be obtained if required, such as detrending or filtering. The type and number of pre-processing operations depend on several factors, such as data quality, type of test or special conditions, which may affect the expected force shape described in Section 2. The goal is to clean and prepare the data in order to ease the next stages.

The selection process is a manual stage intended to remove spurious forces caused by movements different from walking that can typically be found at the beginning and ending of the raw recorded data. The

(a) (b)

Fig. 3. (a) Loadsol® Instrumented insoles and (b) Example of instrumented insoles usage.

final data arrays should contain as many consecutive steps as possible. Otherwise, the developed algorithm and the statistical analysis may not provide meaningful results. This task is usually performed by manually setting a time interval $[t_{min},\ t_{max}]$ during which the relevant data have been recorded and any data falling outside that time interval are discarded.

3.3. Step detection

After the data to be analyzed have been selected, the algorithm starts by identifying the steps taken by each foot. The detection algorithm is applied to both vectors separately and it provides two output vectors per foot: one with the specific samples at which every step starts and another with the samples at which every step finishes. To do so, a flag (true/false) is raised based, on a four-samples window, which is moved sample by sample along the vector under analysis: the window can be said to be inside a step if the flag is set to true, or outside if the flag is set to false. For the ith window $w_i = [w_{i1}, w_{i2}, w_{i3}, w_{i4}]$, with $i > 1$, a step is detected when it meets the following conditions:

- The previous window, w_{i-1}, was outside a step (the flag was set to false)
- Its elements monotonically increase, so $w_{i1} < w_{i2} < w_{i3} < w_{i4}$
- The value of its first sample is above a previously determined threshold ($w_{i1} \geq F_{th}$)

If the three conditions are met, the flag is set to true: w_{i1} is saved as the step's starting sample, and the current and subsequent windows are marked as part of the detected step. Fig. 4(a) shows an example where the flag is set to false at the beginning. The window composed of the samples 1, 2, 3 and 4 does not monotonically increase, so it does not activate the detection flag. The samples 4, 5, 6 and 7 monotonically increase, but the first sample is below the specified threshold. The step is detected, and the flag is set to true, by the window formed by the samples 5, 6, 7 and 8, while sample 5 is saved as the starting one for that step. The window formed by the immediately following samples are said to be inside the step so, even if they monotonically increase and the first sample is above the threshold, a new step is not detected as the first condition is not met.

The end of a step is detected similarly. The ith window $w_i = [w_{i1}, w_{i2}, w_{i3}, w_{i4}]$ must satisfy these conditions in order to set the flag to false:

- The previous window must be inside the step (flag equals true)
- Its elements monotonically decrease, so $w_{i1} > w_{i2} > w_{i3} > w_{i4}$
- The fourth element of the window must be below the threshold, $w_{i4} \leq F_{th}$

When a window satisfies the conditions, the flag is set to false; the fourth element of the window, w_{i4}, which is the one right below F_{th}, is saved as the ending sample. An example is also shown in Fig. 4(b), where the flag is set to true at the beginning. The window composed of the samples 2, 3, 4 and 5 meets the previous criteria, so sample 5 is saved as the last sample of the current step and the flag is set to false. The next windows are considered to be outside the step, so eventually a new step could be detected some samples later.

Note that the threshold value F_{th} is required in order to avoid issues associated with noise around 0 N, mainly caused by sensor noise. The value of this threshold should be higher than the sensor noise, but small enough to lose as few samples as possible at the start and end of the detected steps. In this regard, a value of 20 N is used. Note that losing some extreme points may make the steps look shorter than they actually are and may affect the modeling of the double support phase. In the next section, a way to overcome this issue is proposed.

3.4. Step characterization

Once the steps are detected, they are isolated, scaled, and resampled to ease the subsequent statistical analysis.

3.4.1. Segmentation

The complete set of samples for each step is extracted from the original force vectors using information from the previous section. Each step i may have a different number of samples n_i, resulting in varying durations. Additionally, the n_i samples may not correspond to uniform relative time intervals. Fig. 5 illustrates this: the black continuous line represents the true force at the start of the step, while the blue, red, and green points represent different discretizations. Due to discretization and threshold-based detection algorithm, the initial sample appears slightly after the step onset, varying across discretizations (e.g., 14.070 s for blue, 14.073 s for red, and 14.067 s for green), leading to inaccuracies in step duration and double support times t_{LR} and t_{RL}.

To address this issue, a technique based on extrapolation is proposed. The core concept is to extrapolate the first and last few samples to estimate the time instants at which the step force reaches 0 N, denoted as t_0^* at the beginning and t_{end}^* at the end. The symbol * indicates that these values are approximations. While not exact, this method improves upon assuming the step begins and ends at the first and last recorded samples. The proposed extrapolation uses the first and last four samples and applies a linear extrapolation that reaches 0 N (nonlinear methods might yield curves that fail to cross the horizontal axis). Fig. 5 illustrates this process: the blue points are linearly extrapolated to provide a better estimate of the step start time ($t_0 \approx t_0^* = 14.064$ s)

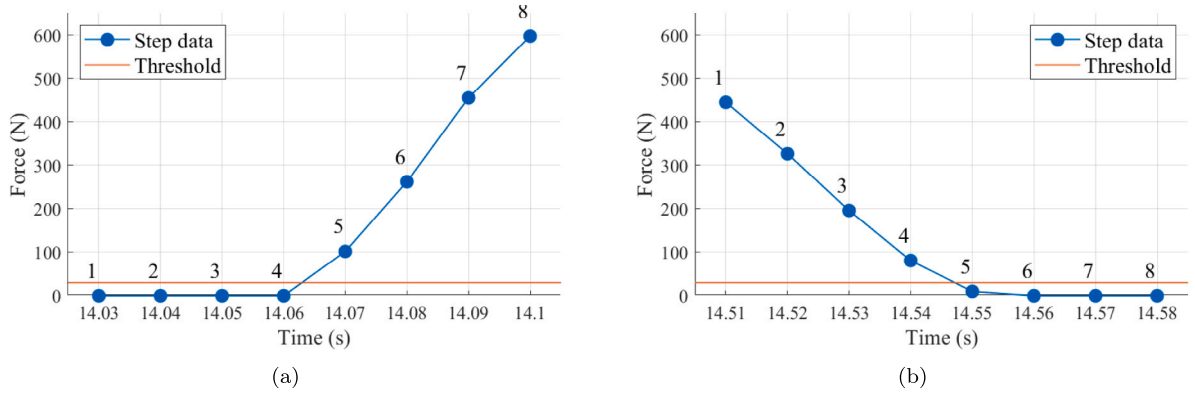


Fig. 4. (a) First samples of a step and (b) last samples of a step with a threshold of $F_{th} = 20$ N (100 S/s).

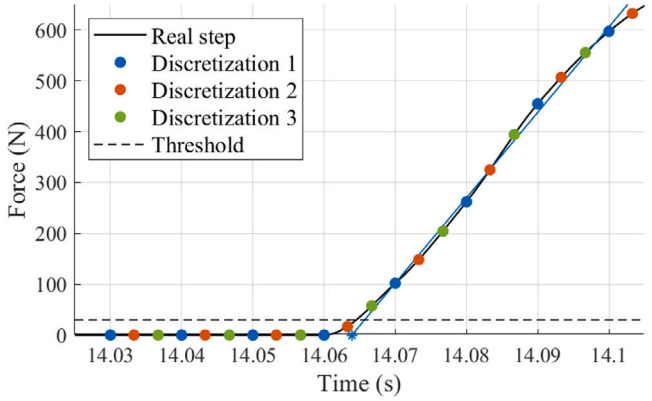


Fig. 5. Various discretizations of the same step (100 S/s) and extrapolation of the blue one to estimate its start.

compared to the first detected sample at 14.07 s. As a result, two artificial samples are added to each step: $(t_0^*, 0)$ and $(t_{end}^*, 0)$. The total duration of the ith step is then estimated as $\delta_i^* = t_{end,i}^* - t_{0,i}^*$.

Before proceeding with the scaling and resampling procedures, an initial selection of steps is needed. As previously mentioned, individual steps differ and, statistically, some steps may be significantly less representative than others. Including all steps in the analysis, regardless of their significance, could result in stochastic models that fail to accurately characterize how a person walks. Therefore, it is crucial to identify and remove less significant steps prior to stochastic characterization. Two criteria are proposed for marking steps for removal: one based on step duration and another based on step shape.

Regarding the duration, which is assumed to be a normally distributed random variable (this assumption will later be validated with real data), outliers are identified using the associated quartiles (Q1 at 25%, Q2 at 50%, and Q3 at 75%) and the interquartile range (IQR = Q3 − Q1). In this study, outliers are defined as steps with durations exceeding 1.5·IQR above the upper quartile Q3 or below the lower quartile Q1. For reference, in the case of symmetrical normal distributions, this corresponds to excluding values outside the range $\pm 2.698\sigma_\delta$ around the mean μ_δ. Notably, this criterion is independent of the specific shape of the step force and can be applied universally, regardless of the type of activity being analyzed (e.g., walking, running, etc.).

Finally, concerning the shape of the step and referring to the general description in Section 2, walking steps typically exhibit two relative maxima and an intermediate local minimum. The second criterion for marking steps for removal involves verifying the presence of these three elements. Numerous algorithms exist for identifying local maxima in numerical data, while the intermediate minimum can be located by finding the maximum of the negative series of data between the two preceding maxima. To be deemed significant, a step must contain exactly these three elements, and each element must not be an outlier of its own distribution; otherwise, it is marked for removal before statistical modeling. This outlier detection method is applicable only when measuring the vertical GRFs of healthy people walking comfortably. For other cases, where different shapes are expected, the detection criteria should be adapted or the outlier removal process omitted.

3.4.2. Step scaling and resampling

Each step has an estimated duration δ_i^* and a different number of samples n_i. Before performing the statistical analysis, it is necessary to standarize all steps to have the same duration and number of samples. Each step is normalized by subtracting the initial time value, t_0^*, from all time values and dividing by the total duration. As a result, all steps start at $\tau = 0$ and end at $\tau = 1$. The scaling factors, $\lambda_{i,L}$ for left and $\lambda_{i,R}$ for right steps, are recorded for stochastic modeling. To recover original durations, the normalization is reversed using the mean values $(\mu_{\lambda,L}, \mu_{\lambda,R})$ and standard deviations $(\sigma_{\lambda,L}, \sigma_{\lambda,R})$ of the assumed normal distributions (validation in Section 4) of each foot individually.

Once steps have the same duration, the sample number issue is addressed through interpolation. By fixing the first and last samples at (0,0) and (1,0), N points are interpolated between the originals. All steps are then resampled with the same number of points $(N_1 = N_2 = \cdots = N)$. Cubic Hermite interpolation, ensuring C^1 continuity, i.e., continuity and interpolation up to the first derivative of the interpolated sample, is used to preserve the step shape. The resulting steps share the same sample count between (0,0) and (1,0).

3.4.3. Stochastic modeling of the steps

With the previous stages completed, the statistical model representing one foot's steps is obtained. Each of the N samples from the m resampled steps is treated as a stochastic variable with m observations (one per step). Assuming all variables are dependent, generating new steps requires using a multivariate random number algorithm to generate all points simultaneously. As described in [26], the model is based on the means of each variable and the covariance matrix. The diagonal elements represent the variances, while the off-diagonal elements show how they relate to each other. Thus, the means of the N variables for each foot, μ_L or μ_R, and the covariance matrices, S_L or S_R, are computed and saved as two N-dimensional stochastic models, one per foot. Separate models are necessary if gait symmetry is not assumed, allowing individual analysis of each foot's GRF. Fig. 6 summarizes the methodology presented in this and the previous subsections.

Finally, the resampled steps can be randomly divided into two sets following this strategy: one subset is used to obtain the statistical model, while the other is reserved to validate the generation of virtual steps afterwards. This division ensures that the model is built independently of the validation process, allowing for an assessment of its accuracy in terms of geometry and shape of the virtual steps, as well as generalization to unseen data. This is done in Section 4.

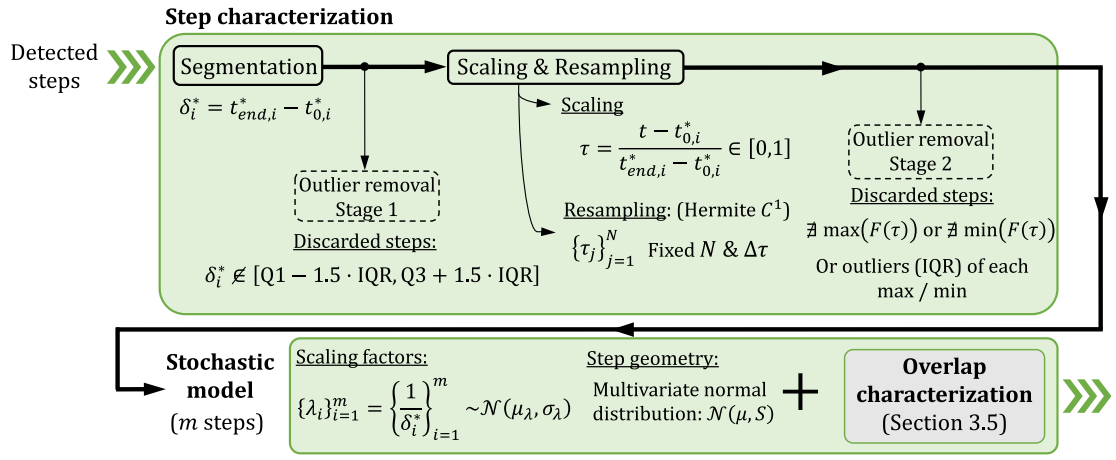


Fig. 6. Overview and summary of the step characterization methodology for each foot.

3.5. Double support characterization

The model obtained after applying the algorithm described in Section 3.4 represents the right or left steps separately, but it does not account for how they relate to each other. After the vector containing the forces is segmented to isolate the steps, the double support times, t_{LR} and t_{RL}, are no longer represented in the data being processed. As stated in Section 2, this variable is crucial if the resultant GRF is of interest, since it is closely related to the maxima of the resultant force that a pedestrian applies to the ground while walking. To start with, the data recorded by each separate insole must be synchronized with each other. This is ensured, if not done automatically by the datalogger, using trigger signals at the beginning of each walk.

The double support time is not constant over time, but a random variable that needs to be characterized. In order to do so, the starting and ending times for each step estimated in Section 3.4 are used to calculate how much time both feet are on the ground, as shown in Fig. 7. This can be done by subtracting the start time, t_0^*, of a certain step taken with one foot from the ending time, t_{end}^*, of the previous step taken with the other foot. Eq. (1) represents more precisely the mathematical definition of the double support times t_{LR} and t_{RL}, where the letters L and R have been appended to the subscripts of t_0^* and t_{end}^* to refer to the left and right foot, respectively.

$$\begin{aligned} t_{LR} &= t_{end,L}^* - t_{0,R}^* \\ t_{RL} &= t_{end,R}^* - t_{0,L}^* \end{aligned} \tag{1}$$

Multiple values for t_{LR} and t_{RL} are computed, and all steps are included in this process, even those marked as outliers in the preceding stages. Since there is a strong relationship between the duration of the double support phase and the main peak of the resultant GRF (P_4 in Fig. 1), a very short overlapping time would result in unrealistically low peaks, while an excessively long overlap time would lead to unrealistically high peaks. To address this issue, a simple yet efficient strategy is incorporated into the methodology.

1. The main peaks of the resultant force are calculated and stored for each double support phase. Both LR and RL phases are identified after labeling overlapping times using Eq. (1)
2. Peaks are analyzed for misdetections (absent values) or outliers (using the IQR method). Overlapping times producing such situations are removed
3. A new IQR-based outlier removal is applied with times associated with atypical peaks (low or high) already excluded

Double support times t_{LR} and t_{RL} are assumed to be properly represented by a normal distribution and their corresponding means (μ_{LR} and μ_{RL}) and standard deviations (σ_{LR} and σ_{RL}) are computed in order to obtain the sought stochastic models.

3.6. Virtual GRF generation

The stochastic model obtained in the previous section is composed of the following:

- Means, $\mu_{\lambda,L}$ and $\mu_{\lambda,R}$, and standard deviations, $\sigma_{\lambda,L}$ and $\sigma_{\lambda,R}$, of the scaling factors of left and right feet (Section 3.4.2)
- Two mean vectors, μ_L and μ_R, and the corresponding covariance matrices, S_L and S_R, that account for the individual steps (left and right separately). The size of this model is N, the number of samples in which the steps have been resampled (Section 3.4.3)
- Means, μ_{LR} and μ_{RL}, and standard deviations, σ_{LR} and σ_{RL}, of the double support times (Section 3.5)

The total number of variables (T_V) that the model comprises of, and the corresponding total number of parameters (T_P), are given in Eq. (2). Each scaling factor and double support time represent a single random variable with two associated parameters (μ) and (σ), leading to a fixed total of 4 variables and 8 parameters. Since the size of each step model is N, an additional $2N$ random variables are introduced (left and right feet). The minimum number of unique parameters associated with these variables is $2N + N(N+1)$, as the covariance matrix is square and symmetric ($N \times N$), and each random variable of the multivariate distribution has a mean (all gathered together in mean vectors of length N). Taking this into account, N has to be chosen in order not to oversize the model, while also ensuring that there is sufficient information available.

$$\begin{aligned} T_V &= 4 + 2N \\ T_P &= N^2 + 3N + 8 \end{aligned} \tag{2}$$

The virtual GRF generation is divided into two stages: step generation and step concatenation. By means of a multivariate normal distribution generation algorithm, a set of $N_{g,L} + N_{g,R}$ steps are first generated, $N_{g,L}$ for the left and $N_{g,R}$ for the right foot. This stage makes use of the mean vectors, μ_L and μ_R, and their corresponding covariance matrices, S_L and S_R. In this work, the algorithm programmed in the MATLAB function `mvnrnd()` is used for the task. Note that, as a consequence of how the stochastic model has been obtained, the virtual steps are normalized in time, so they all cover the range between $\tau = 0$ and $\tau = 1$. To solve this, the minimum common number $N_g = \min(N_{g,L}, N_{g,R})$ of pseudo-random numbers are generated by using the means $\mu_{\lambda,L}$ and $\mu_{\lambda,R}$, and standard deviations, $\sigma_{\lambda,L}$ and $\sigma_{\lambda,R}$, associated to the scaling factors. Normalization is then reversed by using those numbers, one per step, so each step now has a different duration ($t_{L,i}$ for the left steps and $t_{R,i}$ for the right ones).

The resulting steps are then concatenated. The virtual double support time between two consecutive steps (LR and RL) is generated

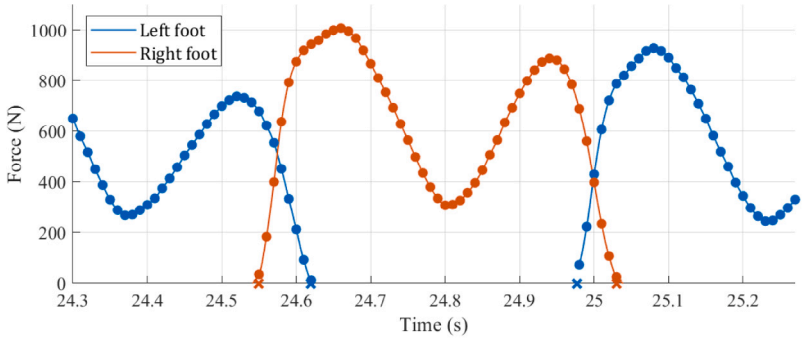


Fig. 7. Example of overlapping time of the double support phase for the two scenarios, LR and RL (100 S/s).

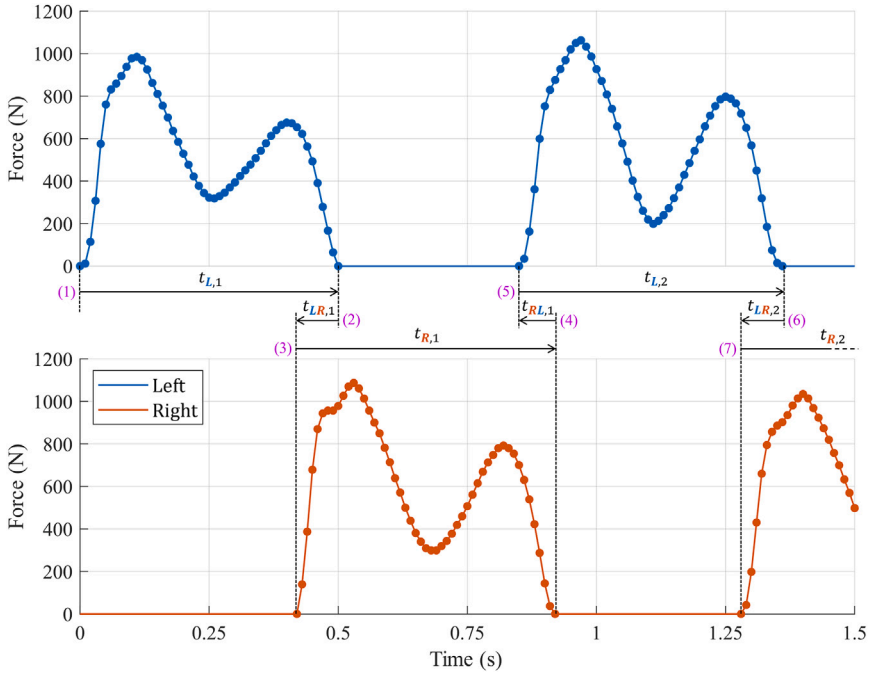


Fig. 8. Example of virtual GRF generation (100 S/s) with the order in which virtual steps are sorted (pink).

through their means, μ_{LR} and μ_{RL}, and standard deviations, σ_{LR} and σ_{RL}. The overlap values are used to assign the starting time for each step based on the time at which the previous step ends. More specifically, as shown in Fig. 8, the first step is assigned to start at $t = 0$. It ends at a certain time, $t = t_{L,1}$. The second step, which is taken by the other foot, is assigned to start at $t = t_{L,1} - t_{LR,1}$, where $t_{LR,1}$ is the first virtual t_{LR} overlapping time. The process is repeated until $2N_g$ steps (N_g of each foot) have been assigned a starting and ending time in a global time axis, common to both left and right feet.

GRFs have uneven time vectors, since each step has a different time increment (as a consequence of reversing their scaling, keeping the same number of samples per step) and the time it takes from the end of one step and the beginning of the next one is notably greater. To fix this, both vectors are resampled simultaneously in order to share a common time axis. It is important to note that the new values interpolated between two consecutive steps made by the same foot need to be zero. These points are associated to the swing phase of the walking process, and the force made by that foot in that situation is zero by definition (Fig. 1).

As a result of the previous process, two new vectors with virtual GRFs are generated. These forces are statistically similar to the ones used to obtain the stochastic model, so they are assumed to be similar to the forces that the pedestrian who generated the original data may induce. Note that the model associated to a specific pedestrian may change due to several factors: walking pace, type of shoes, floor vibration level, accompanying people, etc. One different stochastic model should be available for each different situation, leading to a high amount of information to be handled. In this work, only the possibility of changing the walking pace is illustrated.

4. Application examples

This section is devoted to showing and discussing the results obtained after applying the described methodology to a set of registered steps at different walking paces. With the obtained models, virtual GRFs are generated and compared with the experimental ones to show the goodness of the developed algorithm.

4.1. Measured signals and testing protocol

The data were recorded using the insoles described in Section 3.1 and Fig. 3, worn by a 55-year-old healthy male (93.3 kg, common medical scale, 177 cm in height). The pedestrian walked at four different paces (steps/min): 110, 120, 130, and 140, assisted by a digital portable metronome. Trials were performed on a $L = 100$ m straight, obstacle-free path. The recorded data consist of two force vectors (one per foot)

Table 1
Walking paces and their corresponding average velocities and step lengths in the conducted trials.

Pace (steps/min)	Freq. (Hz)	Velocity (m/s)	Step length (m)
110	1.82	1.19	0.654
120	1.99	1.37	0.688
130	2.15	1.63	0.758
140	2.32	1.77	0.763

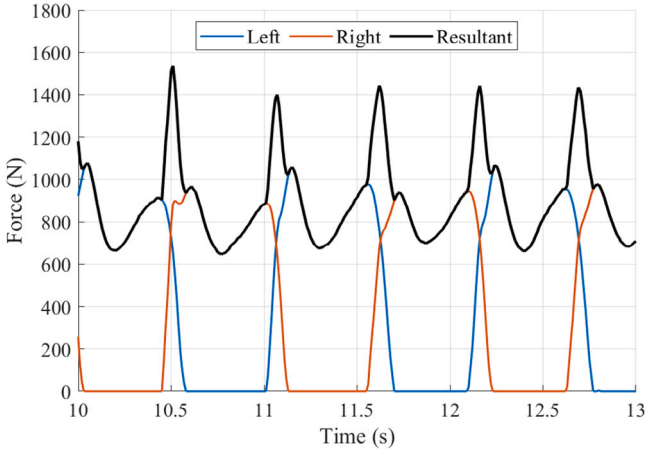


Fig. 9. Example of recorded resultant force (110 steps/min).

sampled at 100 S/s, with consecutive points separated by $\Delta t = 0.01$ s. As the insoles are synchronized, the ith samples from both vectors are assumed to occur simultaneously.

The testing protocol involved the pedestrian wearing the insoles and calibrating them individually for each foot at the beginning of the first trial (110 steps/min) using the previously measured body weight. At the start of each trial, data recording was initiated, and the pedestrian walked the 100 m path. Upon reaching the end, the recording was stopped and the data saved. After a 2 min rest period, the pedestrian performed another trial, increasing the walking pace by 10 steps/min. Additionally, for synchronization purposes, the start and end of each trial were marked with a small trigger (a light static jump) applied simultaneously with both feet. Average velocities and step lengths are presented in Table 1 and were estimated for each walking pace as follows:

1. Average velocity was calculated as $v = L/t$, where t is the total trial duration, while frequencies correspond to the first harmonic (f_0) of the resultant GRF's Fast Fourier Transform
2. Average step length for both feet was computed by obtaining the ratio between estimated walking velocity and walking frequency (v/f_0)

After pre-processing the vertical GRF signals to make a first selection of the relevant data (Section 3.2), a total of 560 steps were detected (see Table 3), with both 76 left and right foot steps at 110 steps/min, followed by 73 left and 72 right at 120 steps/min. At 130 steps/min, 66 left and 67 right steps were logged, and 65 for each foot at the fastest trial (140 steps/min). Since L is fixed at 100 m, the number of steps decreases as the rate increases. Each trial is analyzed and processed individually. Fig. 9 illustrates a detail of the recorded left and right GRFs at 110 steps/min and the corresponding resultant force, where the peaks during the double support phase are well visible.

4.2. Stochastic model

The recorded force vectors are processed as described in Section 3.3. The threshold is set at 20 N, with steps starting when the 4-sample window exceeds this value, and ending when the last sample is slightly below 20 N.

Table 2
Shapiro–Wilk normality test results ($\alpha = 0.05$) for step durations.

Walking pace	p-value		W	
(steps/min)	δ^*_L	δ^*_R	δ^*_L	δ^*_R
110	0.6411	0.3726	0.9862	0.9818
120	0.04883	0.9841	0.9656	0.9939
130	0.2657	0.2181	0.9767	0.9766
140	0.01749	0.1804	0.9531	0.9743

Segmented steps are first extended by extrapolating the first and the last four samples to find the times at which the force equals 0 N. As stated in Section 3.4, this is necessary to improve their duration estimation and better model the double support time. Most steps now start at a negative relative time ($t^*_0 < 0$ s), but this is easily corrected by summing t^*_0 to all the values in the time vector. In any case, it is now easier to estimate the step duration by subtracting the estimated starting time value t^*_0 from the end time value t^*_{end}. In addition, note that any outlier step has not yet been removed. Fig. 10 depicts the steps at 110 steps/min with the duration outliers marked for removal. For example, Fig. 10(a) depicts 4 left steps that last significantly more than the others ($\delta^* > \text{Q3}+1.5 \cdot \text{IQR}$), while one step lasts significantly less ($\delta^* < \text{Q1}-1.5 \cdot \text{IQR}$). In Fig. 10(b), 2 steps are longer than usual and one is shorter. The first step detected in both feet is not plotted for representation purposes, as its t^*_0 makes it a clear outlier.

After removing the duration outliers (first stage of the outlier removal process), histograms of the associated step duration of each foot are shown in Fig. 11. Each random variable is assumed to follow a normal distribution, so this hypothesis is now tested using the Shapiro–Wilk normality test (MATLAB function available at [27]) in order to generate the required pseudo-random values for all the walking paces, not just the one at 110 steps/min used as an example here. The test results are shown in Table 2, by means of the corresponding p-values, test statistics (W), and a standard significance level of $\alpha = 0.05$. Results show that most of the step duration data follow a normal distribution (p-Value $> \alpha$), with some close but bearable exceptions at 120 and 140 steps/min (left steps).

The next modeling stage consists of the step normalization, which transforms the absolute time axis $[t^*_0, t^*_{end}]$ into a relative one $[\tau = 0, \tau_{end} = 1]$. The scaling factor is the inverse of the step duration and it is modeled through its mean ($\mu_{\lambda,L}$ for the left steps and $\mu_{\lambda,R}$ for the right steps) and standard deviation ($\sigma_{\lambda,L}$ and $\sigma_{\lambda,R}$). These values are useful for generating the necessary pseudo-random values to reverse the normalization of virtual steps.

A second outlier removal stage is performed by checking, step by step, the proper presence of two local maxima and intermediate local minimum. Furthermore, the steps whose maxima or minima values are significantly different from the others are also treated as outliers. Fig. 12 shows the identified local maxima and the intermediate local minimum for both left and right feet at 110 steps/min. In all of the proposed trials at different walking paces, all of the steps possess both local maxima and the local minimum, so no step has been removed for that reason. However, some values fall outside the acceptation bounds, which are defined as 1.5·IQR under Q1 and above Q3. Table 3 shows the remaining steps at each walking pace after the outlier removal process. Taking into account this stage, the final normal parameters of the scaling factors (and Shapiro–Wilk test results) are presented in Table 4. Note that normality test results differ from the ones in Table 3 for δ^*, since now the scaling factors associated to the shape outliers have been removed. Except a slight discrepancy for λ_L at 140 steps/min, the whole data set is inside test limits.

Now, the remaining steps, which share a common normalized time axis but differ in the number of samples, are resampled (Hermite interpolation, Section 3.4.2) to make all their points share the same N

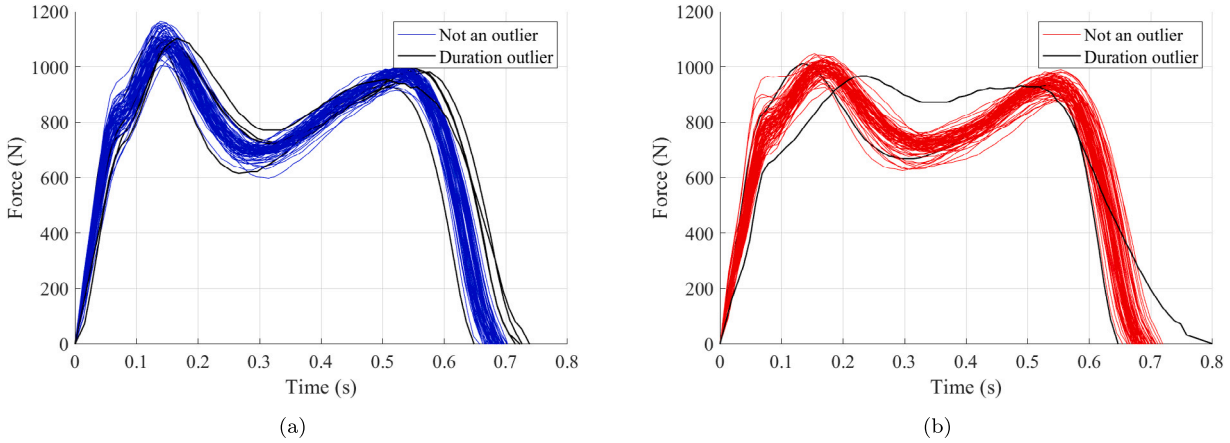


Fig. 10. Steps extended up to the 0 N value and duration outliers (110 steps/min): (a) left steps; (b) right steps.

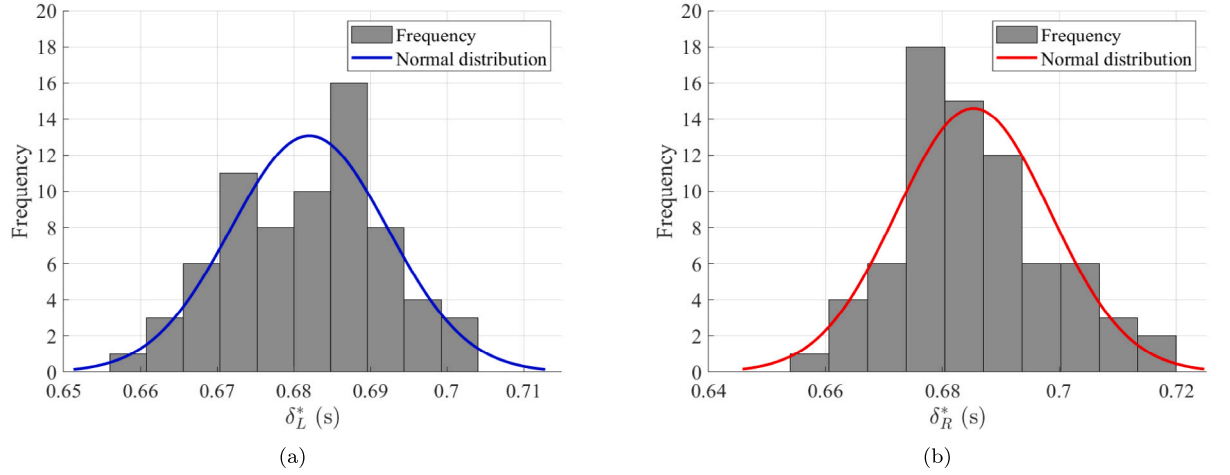


Fig. 11. Histograms and normal distributions of step durations (110 steps/min): (a) left steps; (b) right steps.

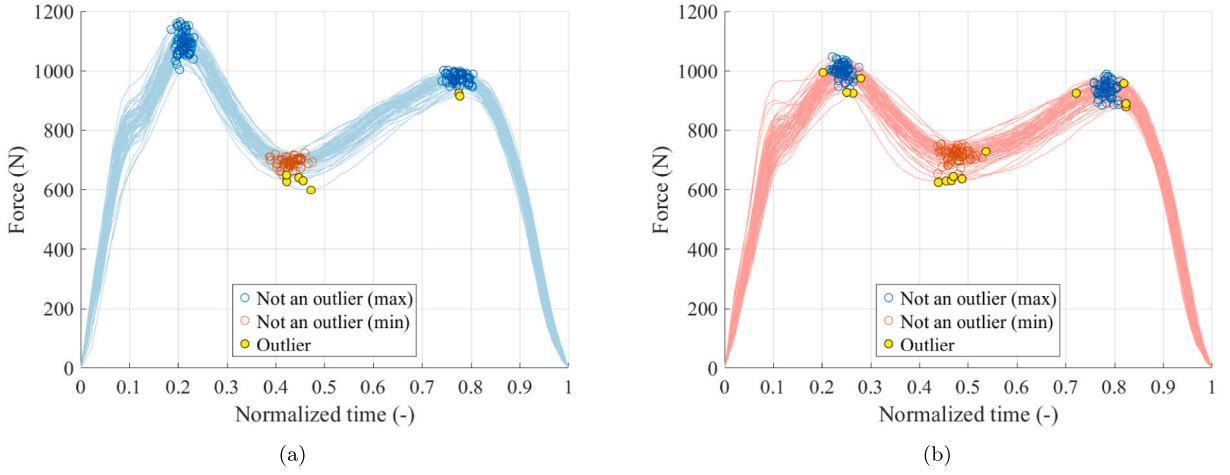


Fig. 12. Detected local maxima and minima with outliers (110 steps/min): (a) left steps; (b) right steps.

timestamps, which are fixed at a total of 50. The latter value has been chosen because of the insoles sampling rate (100 S/s), which implies that all steps originally have approximately less than 70 samples (this number decreases once the walking pace increases), so $N = 50$ makes it a good choice to ensure enough points without oversizing the model in terms of the variables and parameters used (Section 3.6 and Eq. (2)). Table 5 shows a comparison of the different increasing values of N and the corresponding number of variables (T_V) and parameters (T_P). Since the dependence of T_P with N is cuadratic, higher values of N must be avoided for a simpler model.

Each sample, at a certain relative time τ, can be considered as an observation of the random variable associated to that instant. Fig. 13 shows an example of how these variables must be interpreted. In the horizontal plane, the normalized and resampled steps associated

Table 3
Number of steps after each outlier removal stage, shown as left–right (L-R).

Pace (steps/min)	Original	Stage 1	Stage 2
110	76–76	70–73	63–63
120	73–72	71–69	62–65
130	66–67	64–65	59–54
140	65–65	63–63	57–58

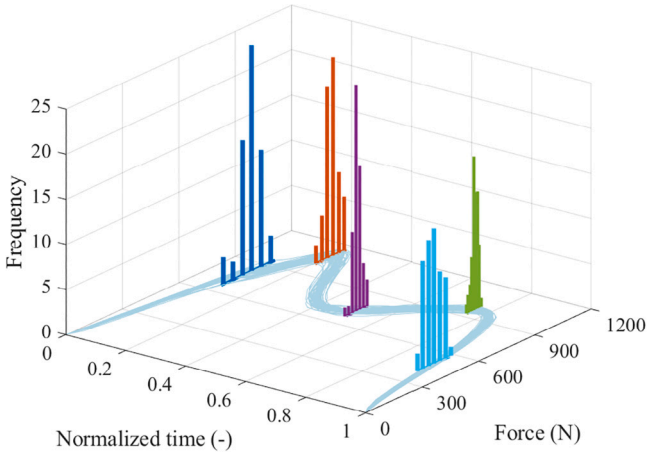


Fig. 13. Example of five distributions at five normalized time instants associated to the left steps (110 steps/min).

to the left foot (110 steps/min) are plotted. For an example of five different relative time instants of the whole N points, the associated histograms are plotted to show how each set of samples can be modeled by means of a normal distribution, with their own mean and standard deviation values. Furthermore, the mean, median and quartiles Q1 and Q3 are computed. Fig. 14 shows these values for the left steps at 110 steps/min, where now only approximately 70% of the total steps after the full outlier removal are used to generate the stochastic model, so shape can be graphically tested later with the rest of the unseen step data (as stated in Section 3.4.3). This is replicated for each walking pace. The mean and median values are very close to each other, a sign of the distribution symmetry. In addition, both quartiles Q1 and Q3 are relatively close to the mean and median values, indicating narrow distributions, something that can also be appreciated in Fig. 13.

As mentioned in Sections 3.4 and 3.6, the entire set of N variables that describe the steps of one foot is treated as a multivariate normal distribution. This means that each variable is correlated with the others, and when generating random steps, all the samples must be generated together. If virtual steps were generated sample by sample, unrealistic noise would occur because each sample could deviate from its mean independently of neighboring samples. Therefore, generating all the samples simultaneously avoids this issue. To fully describe the multivariate normal distribution, the set of mean values for each variable must be obtained, and the covariance matrix, rather than standard deviations or variances, needs to be computed. This matrix is square, with dimensions equal to the number of samples in the resampled normalized steps, N. Fig. 15 checks the normality hypothesis for each of the N variables, foot and walking pace, by means of the corresponding p-values after testing with the Shapiro–Wilk normality test. The majority of the variables are found to follow a normal distribution (p above α).

Now, Fig. 16 shows representations of the covariance matrices, S_L and S_R, associated to the left and right feet of all the examples, where each pixel is associated to a matrix value. Red represents strong dependencies between variables, something that typically occurs during the first (lower left corner of Figs. 16(a) to 16(d)) and ending samples of the step (upper right corner of Figs. 16(c) and 16(d)). As can be

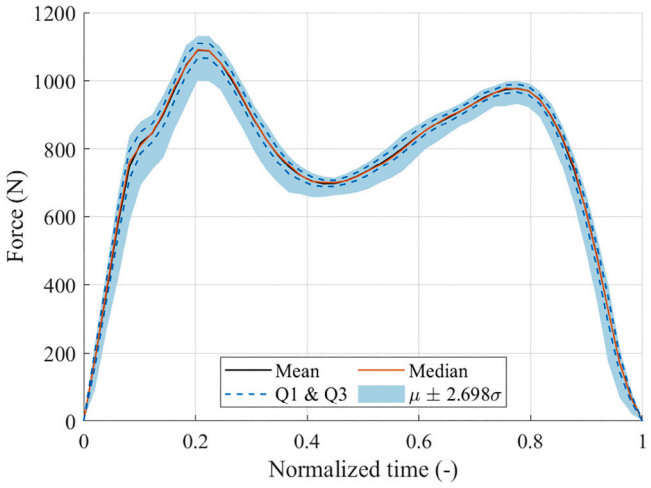


Fig. 14. Statistical description of the normalized and resampled left steps (110 steps/min).

seen, differences arise between both feet and the different walking paces, which allows us to conclude that there exists differences between both feet of this specific pedestrian, since the GRF shape is not the same for both. To give additional evidence of this, Fig. 17 depicts a graphic comparison between the means of the model, showing higher peaks and deeper valleys once the walking pace increases, something that is directly related to the covariance scale, which is different for each trial. Note that the proposed method is fully compliant with these differences.

The presented mean value vectors and covariance matrices constitute a complete model of a multivariate normal distribution that can be used to generate virtual steps, as done in Section 4.3. In order to generate complete virtual GRFs and compute their sum, the double support time needs to be characterized as a final task for the stochastic modeling, as stated in Section 3.5. This is done by calculating the time both feet are on the ground simultaneously for each pair of left–right (and right-left) steps. For each phase, the mean and standard deviation are computed to form the model for generating pseudo-random values after outliers are removed, including those associated with the main peaks of the GRF. Fig. 18 shows example histograms of the double support phase durations (110 steps/min), which are modeled as normally distributed. As before, the Shapiro–Wilk normality test results (for t_{LR} and t_{RL}) are provided in Table 6, where means and standard deviations are also indicated. Once again, results show that most data exhibit behavior consistent with normality, though some deviations occur (at 130 steps/min, t_{RL} shows significant deviation, $p = 0.001221$ and $W = 0.9281$). Despite this, and along the same lines as the previous random variables, normal behavior across different conditions is found, suggesting that the data can generally be modeled by normal distributions.

4.3. Virtual forces

The stochastic models obtained in the previous section are now used to generate new, virtual steps and GRFs in order to compare them to the experimental ones. First, virtual steps at each walking pace are randomly generated by using a multivariate algorithm like the one programmed in the `mvnrnd()` function, using the calculated means and covariance matrices. The number of normalized virtual steps differ for each walking pace, since a different number of experimental steps were reserved for testing their shape (70% for modeling, rest for graphical comparison). So, the number of virtual steps corresponds to approximately 30% of the original remaining, $N_{g,L}$ and $N_{g,R}$. This is depicted in Figs. 19(a) to 19(h).

Table 4
Shapiro–Wilk test results ($\alpha = 0.05$) for scaling factors and normal distribution parameters.

Walking pace	p-value		W		μ_λ		σ_λ	
(steps/min)	λ_L	λ_R	λ_L	λ_R	λ_L	λ_R	λ_L	λ_R
110	0.6159	0.4356	0.9846	0.9809	1.464	1.457	0.02026	0.02558
120	0.1552	0.7756	0.9713	0.9899	1.603	1.596	0.02601	0.02300
130	0.2251	0.2344	0.9735	0.9735	1.762	1.761	0.02649	0.02442
140	0.0151	0.2966	0.9460	0.9758	1.896	1.901	0.04055	0.04011

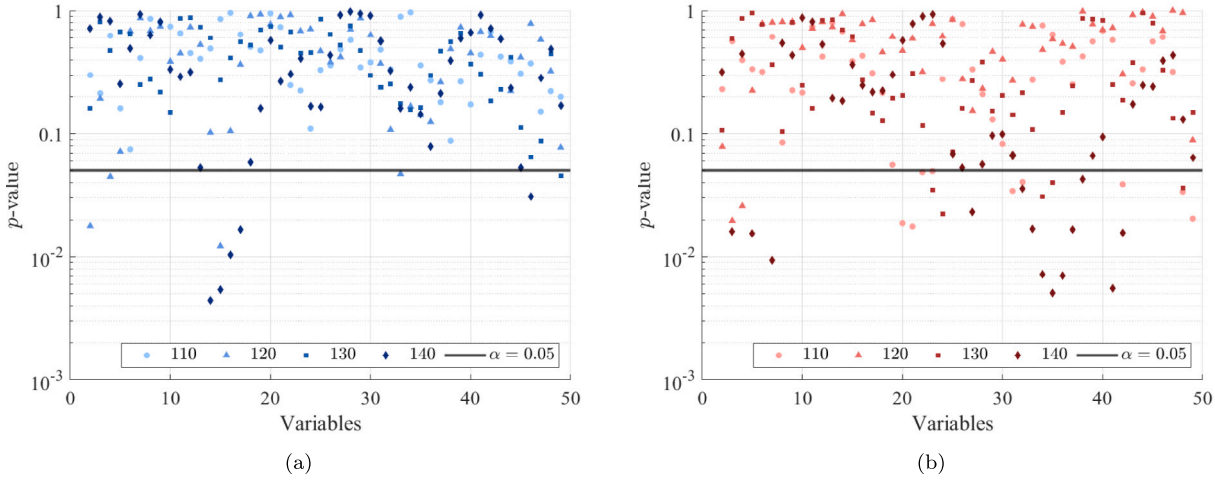


Fig. 15. Shapiro–Wilk test results ($\alpha = 0.05$) for step modeling variables ($N = 50$): (a) left steps; (b) right steps.

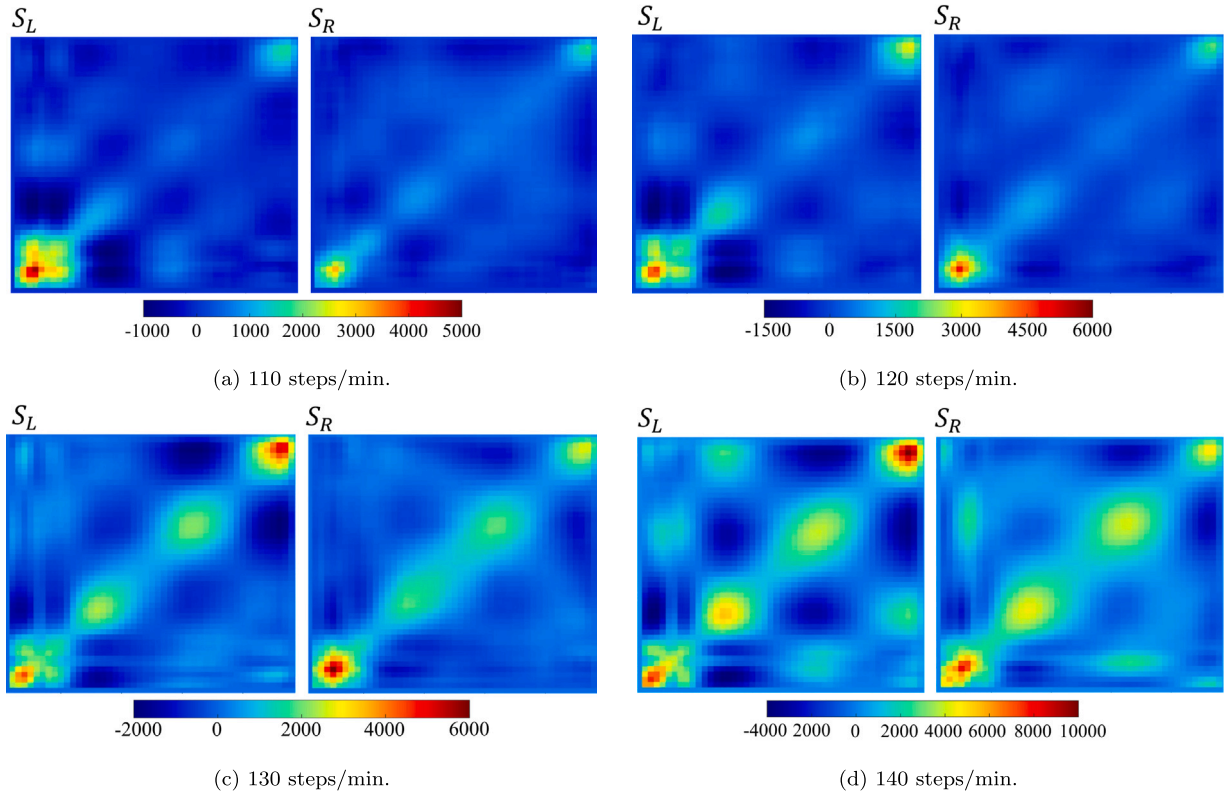


Fig. 16. Covariance matrices for each foot and walking pace ($N = 50$ samples/step, ≈70% of the steps).

As can be seen, the set of new steps are all different from each other, but resemble the original ones in the overall shape, including the slight slope changes at the beginning, mainly caused by the heel striking the floor. Now, scaling of the normalized virtual steps set is possible by reversing the generated scaling factors from their normal distributions, so virtual step durations are achieved for both left and right feet. Additionally, concatenation of $2N_g$ steps ($N_g = \min(N_{g,L}, N_{g,R})$) is made by using two sets of double support time values (left–right t_{LR} and right-left t_{RL}). The resulting GRFs are summed and the resultant force is obtained. In Fig. 20, the virtual force is compared with the

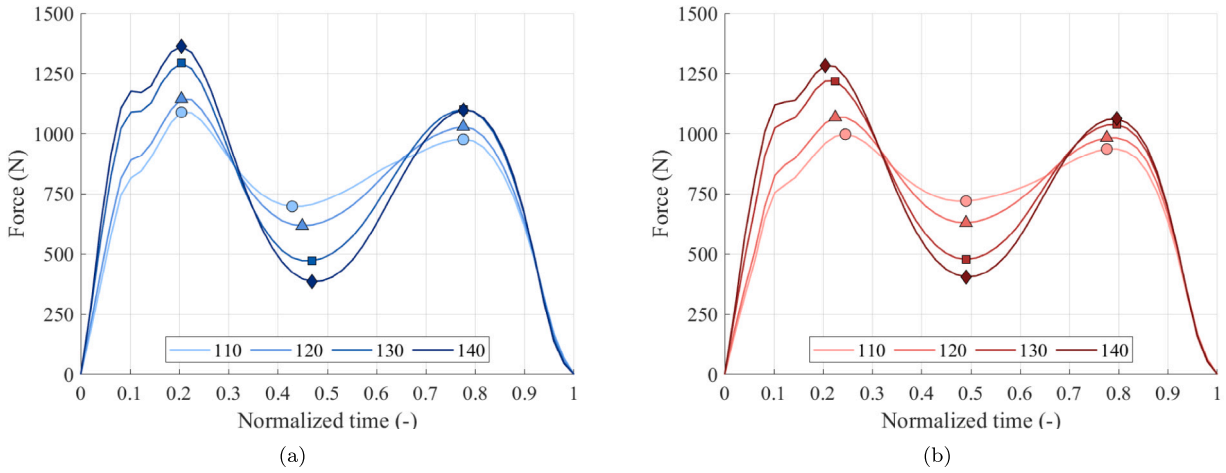


Fig. 17. Comparison between step model means for each walking pace: (a) left steps (μ_L); (b) right steps (μ_R).

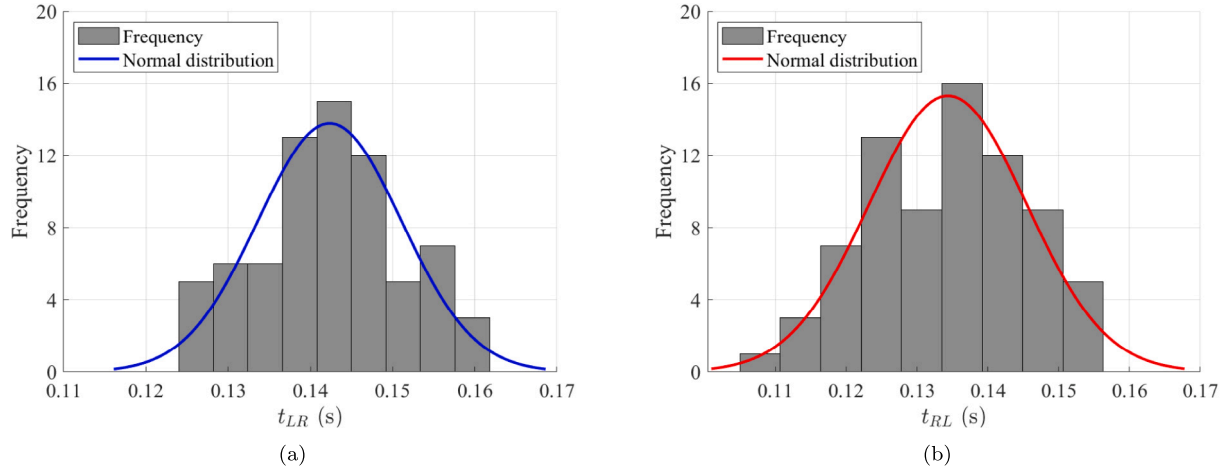


Fig. 18. Histograms and normal distributions of the double support times (110 steps/min): (a) t_{LR}; (b) t_{RL}.

Table 5
Increase (%) in T_V and T_P for different values of N, with 50 as a reference.

N	T_V	$\frac{T_V - T_{V,50}}{T_{V,50}}$ (%)	T_P	$\frac{T_P - T_{P,50}}{T_{P,50}}$ (%)
50	104	–	2658	–
100	204	96.15	10 308	287.81
200	404	288.46	40 608	1427.77

experimental one for each walking pace, where virtual GRFs are finally adjusted to the original insoles sampling rate (100 S/s). Only a 5 s zoom is shown for clarity once both time signals have been phase-aligned. Additionally, the differences presented before while creating the stochastic models are also present here, since the resultant GRF peaks do not show the same behavior for the two possible double support phases, and the virtual forces account for this.

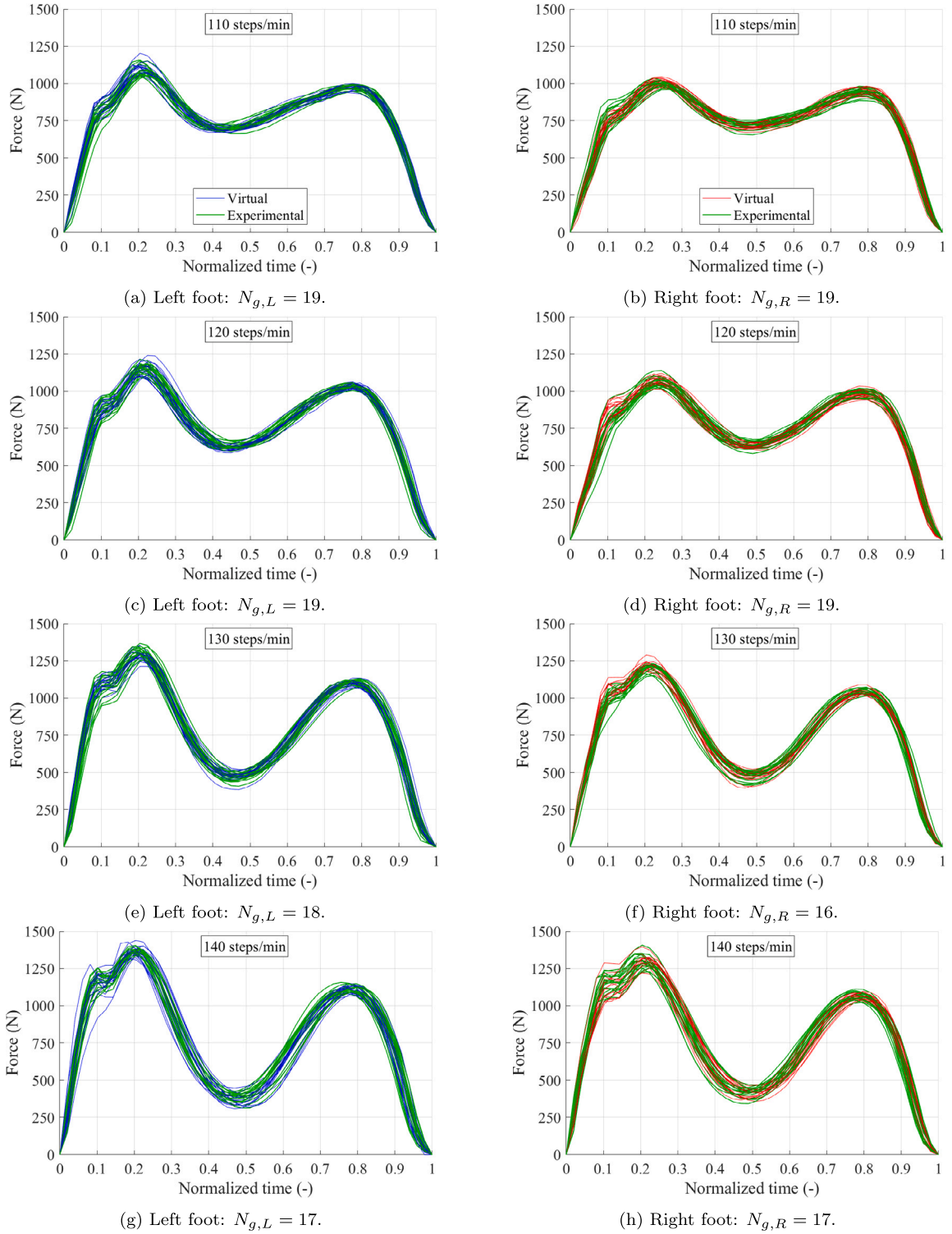
Time-domain virtual signals resemble the pattern presented in Fig. 9. Fig. 21 shows the corresponding Fourier amplitude spectra, where several prominent peaks can be seen at the fundamental walking frequency f_0 and its harmonics ($2f_0$, $3f_0$, $4f_0$, etc.). Finally, Table 7 shows a numeric comparison between peak frequencies obtained from both virtual and experimental transforms at each walking pace, showing a good match between both with the majority of the errors below 2%.

5. Conclusions

This paper proposes a time-domain method for characterizing the forces exerted by a pedestrian while walking, focusing on the vertical component, but adaptable to other components and activities. Consequently, an efficient stochastic, data-driven model has been developed from experimentally measured steps without using Fourier decomposition or Gaussian approximations. The method/algorithm detects and characterizes steps from the raw force data vectors of each foot individually, deriving mean vectors and covariance matrices for each foot, assuming a multivariate normal distribution. This enables

Table 6
Shapiro–Wilk test results ($\alpha = 0.05$) for double support times and normal distribution parameters.

Walking pace	p-value		W		μ_{DS}		σ_{DS} $(\times 10^{-3})$	
(steps/min)	t_{LR}	t_{RL}	t_{LR}	t_{RL}	t_{LR}	t_{RL}	t_{LR}	t_{RL}
110	0.5857	0.4936	0.9856	0.9845	0.1424	0.1345	8.755	11.11
120	0.9797	0.3579	0.9937	0.9806	0.1282	0.1225	9.001	8.954
130	0.4171	0.001221	0.9800	0.9281	0.1113	0.1018	6.621	6.312
140	0.1065	0.03666	0.9668	0.9572	0.0978	0.09791	8.822	9.063



(a) Left foot: $N_{g,L} = 19$. (b) Right foot: $N_{g,R} = 19$.

(c) Left foot: $N_{g,L} = 19$. (d) Right foot: $N_{g,R} = 19$.

(e) Left foot: $N_{g,L} = 18$. (f) Right foot: $N_{g,R} = 16$.

(g) Left foot: $N_{g,L} = 17$. (h) Right foot: $N_{g,R} = 17$.

Fig. 19. Graphical shape comparison between virtual steps and the remaining experimental ones.

Table 7
Peak frequencies of both experimental and virtual GRFs for each walking pace in steps/min.

	Experimental (Hz)				Virtual (Hz)				Error (%)			
	110	120	130	140	110	120	130	140	110	120	130	140
f_0	1.82	1.99	2.15	2.32	1.82	1.99	2.15	2.32	<0.01			
$2f_0$	3.69	3.97	4.36	4.63	3.64	3.97	4.29	4.63	1.36	<0.01	1.61	<0.01
$3f_0$	5.51	6.01	6.51	7.01	5.51	6.01	6.57	7.01	<0.01	<0.01	0.922	<0.01
$4f_0$	7.38	8.00	8.72	9.33	7.33	8.00	8.72	9.26	0.682	<0.01	<0.01	0.750

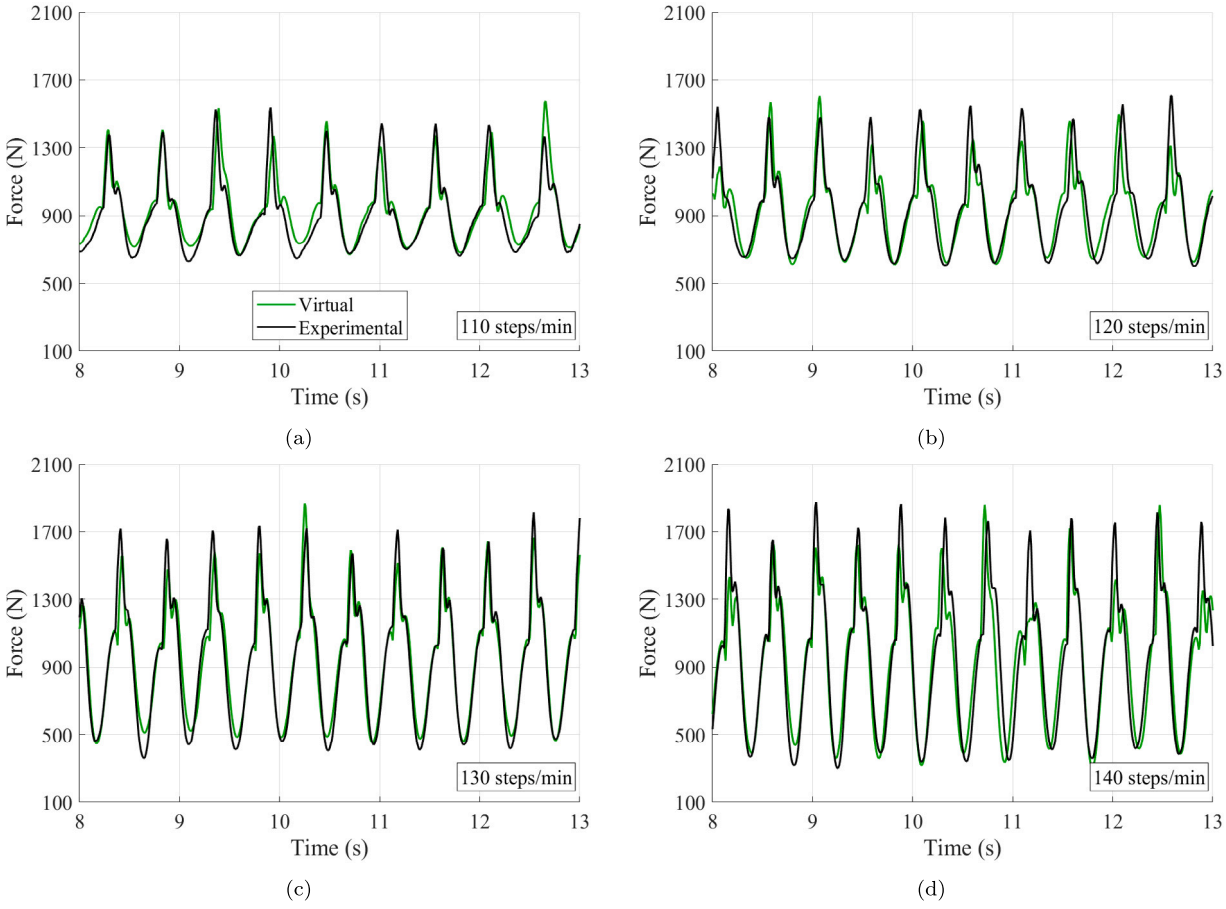


Fig. 20. Comparison between the experimental vertical resultant GRFs and the virtual ones (100 S/s).

the generation of virtual random steps through a purely data-driven stochastic procedure. Mean values and standard deviations represent step scaling and the double support phase as univariate normal distributions, enabling the generation of virtual resultant GRFs. Examples of applications at different walking paces have been provided, comparing the original data with the generated steps. For reproducibility purposes, the data and the MATLAB code have been made available at [25].

The model is composed of two parts: the stochastic model of both separate feet, required to generate the virtual GRFs made with each one; and the stochastic model associated to the parameters that describe the double support phase, necessary to appropriately concatenate the virtual steps and obtain virtual GRFs. One of the main advantages is that the algorithm analyzes each foot individually prior to step concatenation, something useful when forces differ across feet. Although it has been developed in the context of a healthy pedestrian walking at different paces, this methodology could be adapted to other situations by modifying the step detection and the outlier removal conditions. Other, not only vertical GRFs, could be modeled and analyzed, corresponding to steps recorded while running, jumping, or performing any activity that entails a near periodic action of the human locomotion activities. Biomechanics, sports technology, physiotherapy (specifically in locomotion disease diagnosis) and structural engineering – including human–structure interaction – are fields of study in which the present methodology could be applied in the future.

In developing the stochastic models, normal distributions have been used to model the random variables involved. While some deviations from normality have been observed after conducting normality tests, the majority of the random variables follow normal distributions, allowing the results obtained to remain robust and reliable. However, it is important to acknowledge certain limitations. First, the accuracy of the instrumented insoles used in data collection could be improved as stated in the literature. Second, since normality checking relies on the Shapiro–Wilk test conducted only with the present data, a larger dataset could further reinforce this assumption. Additionally, some parameters, such as the choice of $N = 50$ in step resampling, could potentially be refined using proper optimization criteria. As a final remark, please note that the stochastic model may differ from the one presented in this work if the data are gathered in different conditions. The pedestrian may also wear a different footwear, and their biometric parameters (height, weight, etc.) may differ. However, although different from the results presented here, the resulting stochastic model will accurately represent the forces exerted by that pedestrian under those conditions, and the forces generated by that model will only reflect that scenario.

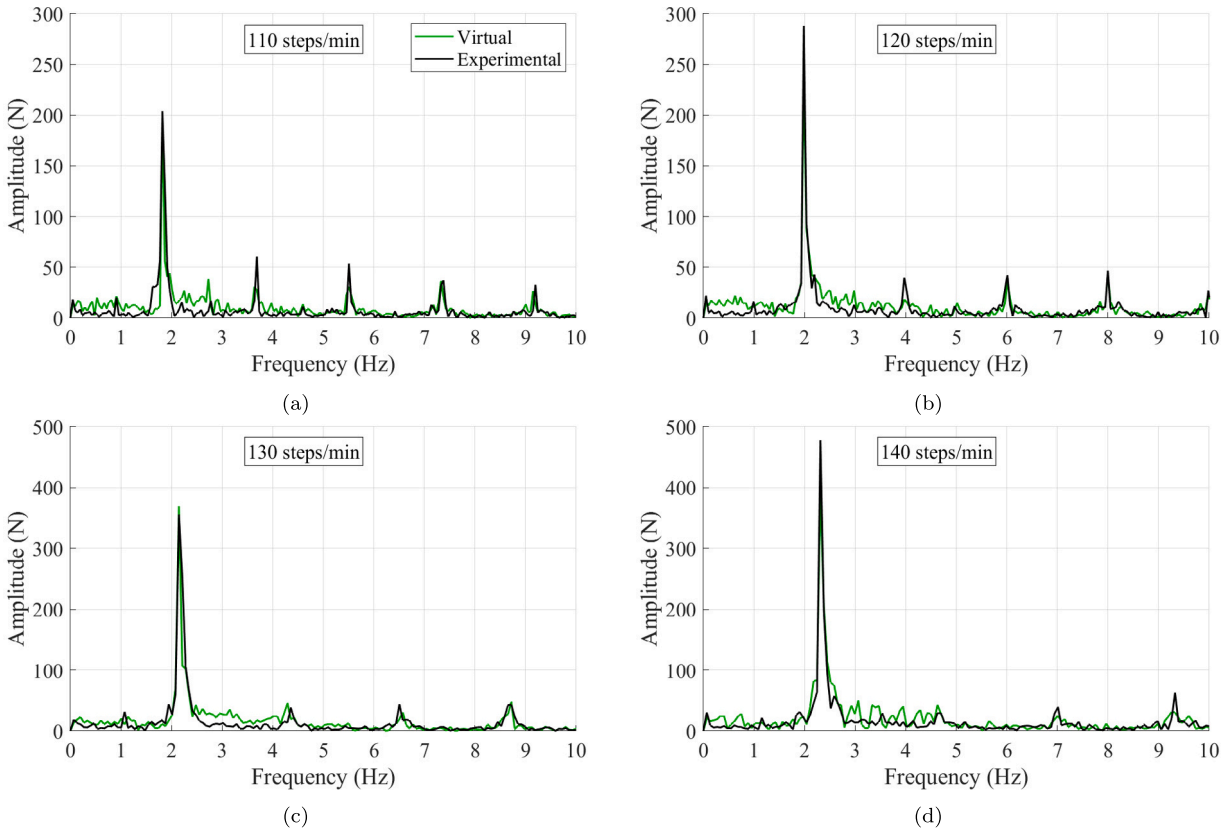


Fig. 21. Fourier amplitude spectra of both the experimental and virtual GRFs in Fig. 20.

In other words, the validity of the obtained model is restrained to the conditions under which the experimental data are registered.

CRediT authorship contribution statement

Alvaro Magdaleno: Writing – original draft, Software, Resources, Formal analysis, Data curation. **José María García-Terán:** Writing – review & editing, Visualization, Methodology, Formal analysis, Conceptualization. **César Peláez-Rodríguez:** Validation, Formal analysis. **Guillermo Fernández:** Writing – review & editing, Software, Investigation. **Antolin Lorenzana:** Supervision, Resources, Project administration.

Funding sources

Funding: This work was supported by Spanish State Research Agency (AEI) and FEDER "ERDF A way of making Europe" (MICIU/AEI/10.13039/501100011033) [grant number PID2022-140117NB-I00]; and NextGenerationEU "InvestigO Program" [grant number CP23-174].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

A Mendeley Data Link has been provided in the manuscript file. This link gives access to the dataset and code and will be freely available.

References

[1] V. Racic, A. Pavic, J.M. Brownjohn, Experimental identification and analytical modelling of human walking forces: Literature review, 2009, http://dx.doi.org/10.1016/j.jsv.2009.04.020.

[2] Y. Ariki, S.H. Hyon, J. Morimoto, Extraction of primitive representation from captured human movements and measured ground reaction force to generate physically consistent imitated behaviors, Neural Netw. 40 (2013) 32–43, http://dx.doi.org/10.1016/j.neunet.2013.01.002.

[3] N.A. Bates, K.R. Ford, G.D. Myer, T.E. Hewett, Impact differences in ground reaction force and center of mass between the first and second landing phases of a drop vertical jump and their implications for injury risk assessment, J. Biomech. 46 (7) (2013) 1237–1241, http://dx.doi.org/10.1016/j.jbiomech.2013.02.024.

[4] C.W. Lin, T.C. Wen, F. Setiawan, Evaluation of vertical ground reaction forces pattern visualization in neurodegenerative diseases identification using deep learning and recurrence plot image feature extraction, Sensors 20 (14) (2020) 1–22, http://dx.doi.org/10.3390/s20143857.

[5] V. Racic, J.M.W. Brownjohn, Stochastic model of near-periodic vertical loads due to humans walking, Adv. Eng. Inform. 25 (2) (2011) 259–275, http://dx.doi.org/10.1016/j.aei.2010.07.004.

[6] V. Racic, J.B. Morin, Data-driven modelling of vertical dynamic excitation of bridges induced by people running, Mech. Syst. Signal Process. 43 (1–2) (2014) 153–170, http://dx.doi.org/10.1016/j.ymssp.2013.10.006.

[7] V. Racic, J. Chen, Data-driven generator of stochastic dynamic loading due to people bouncing, Comput. Struct. 158 (2015) 240–250, http://dx.doi.org/10.1016/j.compstruc.2015.04.013.

[8] F. Pancaldi, E. Bassoli, M. Milani, L. Vincenzi, A statistical approach for modeling individual vertical walking forces, Appl. Sci. 11 (21) (2021) http://dx.doi.org/10.3390/app112110207.

[9] A.E. Peters, V. Racic, S. Živanović, J. Orr, Fourier series approximation of vertical walking force-time history through frequentist and bayesian inference, Vibration 5 (4) (2022) 883–913, http://dx.doi.org/10.3390/vibration5040052.

[10] E. Ahmadi, C. Caprani, S. Živanović, A. Heidarpour, Vertical ground reaction forces on rigid and vibrating surfaces for vibration serviceability assessment of structures, Eng. Struct. 172 (2018) 723–738, http://dx.doi.org/10.1016/j.engstruct.2018.06.059.

[11] M. Cacho-Pérez, A. Lorenzana, Walking model to simulate interaction effects between pedestrians and lively structures, J. Eng. Mech. 143 (9) (2017) 04017109, http://dx.doi.org/10.1061/(ASCE)EM.1943-7889.0001326.

[12] B. Lin, Q. Zhang, F. Fan, S. Shen, Reproducing vertical human walking loads on rigid level surfaces with a damped bipedal inverted pendulum, Struct. 33 (2021) 1789–1801, http://dx.doi.org/10.1016/j.istruc.2021.05.048.

[13] P. Kumar, A. Kumar, V. Racic, S. Erlicher, Modelling vertical human walking forces using self-sustained oscillator, Mech. Syst. Signal Process. 99 (2018) 345–363, http://dx.doi.org/10.1016/j.ymssp.2017.06.014.

[14] H. Wang, J. Chen, J.M. Brownjohn, Parameter identification of pedestrian's spring-mass-damper model by ground reaction force records through a particle filter approach, J. Sound Vib. 411 (2017) 409–421, http://dx.doi.org/10.1016/j.jsv.2017.09.020.

[15] T. Koshio, N. Haraguchi, T. Takahashi, Y. Hara, K. Hase, Estimation of ground reaction forces during sports movements by sensor fusion from inertial measurement units with 3D forward dynamics model, Sensors 24 (2024) http://dx.doi.org/10.3390/s24092706.

[16] C. Peláez-Rodríguez, A. Magdaleno, S. Salcedo-Sanz, A. Lorenzana, Human-induced force reconstruction using a non-linear electrodynamic shaker applying an iterative neural network algorithm, Bull. Pol. Acad. Sci. Tech. Sci. 71 (3) (2023) e144615, http://dx.doi.org/10.24425/bpasts.2023.144615.

[17] S.E. Oh, A. Choi, J.H. Mun, Prediction of ground reaction forces during gait based on kinematics and a neural network model, J. Biomech. 46 (14) (2013) 2372–2380, http://dx.doi.org/10.1016/j.jbiomech.2013.07.036.

[18] J. Nilsson, A. Thorstensson, Ground reaction forces at different speeds of human walking and running, Acta Physiol. Scand. 136 (2) (1989) 217–227, http://dx.doi.org/10.1111/j.1748-1716.1989.tb08655.x.

[19] Z. Han, J.M. Brownjohn, J. Chen, Structural modal testing using a human actuator, Eng. Struct. 221 (2020) http://dx.doi.org/10.1016/j.engstruct.2020.111113.

[20] M. García-Diéguez, V. Racic, J.L. Zapico-Valle, Complete statistical approach to modelling variable pedestrian forces induced on rigid surfaces, Mech. Syst. Signal Process. 159 (2021) http://dx.doi.org/10.1016/j.ymssp.2021.107800.

[21] K. Van Nimmen, G. Lombaert, I. Jonkers, G. De Roeck, P. Van Den Broeck, Characterisation of walking loads by 3D inertial motion tracking, J. Sound Vib. 333 (20) (2014) 5212–5226, http://dx.doi.org/10.1016/j.jsv.2014.05.022.

[22] A.M. Tahir, M.E. Chowdhury, A. Khandakar, S. Al-Hamouz, M. Abdalla, S. Awadallah, M.B.I. Reaz, N. Al-Emadi, A systematic approach to the design and characterization of a smart insole for detecting vertical ground reaction force (vGRF) in gait analysis, Sensors 20 (4) (2020) http://dx.doi.org/10.3390/s20040957.

[23] W. Seiberl, E. Jensen, J. Merker, M. Leitel, A. Schwirtz, Accuracy and precision of loadsol® insole force-sensors for the quantification of ground reaction force-based biomechanical running parameters, Eur. J. Sport. Sci. 18 (2018) 1100–1109, http://dx.doi.org/10.1080/17461391.2018.1477993.

[24] K.E. Renner, D.B. Williams, R.M. Queen, The reliability and validity of the loadsol® under various walking and running conditions, Sensors 19 (2) (2019) http://dx.doi.org/10.3390/s19020265.

[25] A. Magdaleno, J.M. Garcia-Teran, C. Pelaez-Rodriguez, G. Fernandez, A. Lorenzana, Generating vertical ground reaction forces using a stochastic data-driven model for pedestrian walking, Mendeley Data (2024) http://dx.doi.org/10.17632/9twj3w3dsm.1.

[26] Y.L. Tong, Fundamental properties and sampling distributions of the multivariate normal distribution, Multivar. Norm. Distrib., Springer, New York, NY, 1990, pp. 23–61, http://dx.doi.org/10.1007/978-1-4613-9655-0.

[27] A. BenSaïda, Shapiro-wilk and shapiro-francia normality tests, 2024, Version 1.1.0.0, October 10th URL https://tinyurl.com/3secru59.