# A SIMPLE OUTLIER DETECTION METHOD FOR INTRA-SUBJECT TIME-SERIES DATA

**David R. Mullineaux[1] and Gareth Irwin[2]**

**School of Sport and Exercise Science, University of Lincoln, UK[1]**
**Cardiff School of Sport, Cardiff Metropolitan University, Cardiff, UK[2]**

Removal of outliers assists in improving the statistical representations of the general finding. Currently no simple method is advocated for detecting outliers in time-series data obtained in biomechanics. The aim was to demonstrate a 2-stage method for detecting outliers. The test data were the ankle and knee angles for the strides (n=41±2.8) from treadmill running (n=6). Stage 1 was an outlier detection of >±3.3SD from the mean at each time-point, and removing any stride with an outlier. Stage 2, with padding of k=3 points and mean-detrending, was a moving window SD for all strides across ±k data points, and removing strides with any point >±2.58SD. After removal of 5.2±3 (stage 1) and 2.0±1.4 (stage 2) strides, the mean was unchanged and the SD reduced ($p<0.05$). The method is simple and effective in removing outliers in intra-subject time-series data.

**KEY WORDS:** biomechanics, kinematic, outlier, statistic, variability.

**INTRODUCTION:** In data there are often outlier values. In some instances these are real values, and analysing these individual values could be worth exploring to assess if these 'freak' values are a rare event that are optimal or detrimental to performance. In contrast, these outliers can be errors from either subject or experimental sources. Regardless of whether the outliers are 'real' or 'error' they can have a substantial effect on statistical analyses of the data. As both descriptive and inferential statistics are typically used to summarise the results then it is common to remove outliers so these statistics better portray the general finding.

In essence an outlier increases the variability in the data. As such, applications of 'modification' techniques such as transformations, rectification or time-warping (Helwig *et al*., 2011) and offset normalisations (Mullineaux *et al*., 2004) will reduce variability. This will typically lead to fewer outliers. The more complicated these modifications, however, reduces both the external and ecological validity of the data, and makes comparison between studies increasingly more difficult as the analyses and data treatments diverge in similarity. Prior to applying any modification techniques it can be valuable to remove outliers first.

There are many techniques for identifying outliers, and in a review of these it is concluded the choice is dependent on which is most suitable for the problem (Hodge & Austin, 2004). For one-dimensional (1D) data some simple outlier definitions include ±3SD (standard deviations) away from the mean, ±3IQR (inter-quartile ranges) away from the 75th and 25th percentiles (i.e. away from the top and bottom of the box on a boxplot) and ±2.5x1.48MAD (median absolute differences) away from the median.

As it is common in biomechanics to collect time-series data, different techniques to account for the two-dimensions (2D) or influence of data over a range of time-points are required. From areas including economics, techniques vary from simpler (e.g. moving window SD; Brownlees & Gallo, 2006) to more complex ones (e.g. autoregressive models; Kaioti & Caroni, 2004). The use of these in biomechanics would need adapting to cater for the situation, such as, for multiple trials from a single subject. These repetitions of trials collectively have the potential to provide a criterion on which to assess deviations to identify outliers.

The use of a criterion provides the opportunity to use a simpler approach involving a combination of one- and two-dimensional techniques to identify outliers. Improving the criterion is important, hence a first stage would be required for the detection and removal of 1D outliers (e.g. ±3.3SD). A second stage of a 2D outlier detection technique could be employed to take account of the influence of a range of points on each other (e.g. ±2.58 moving window SD). As the two-stages, and particularly the criterion time-series, can be susceptible to individual variability, such a technique would only be suitable for intra-subject trials. Hence, the purpose of this study is to propose a simple two-stage method of detecting and removing time-series outliers in intra-subject data.

**METHODS**: The test data were obtained for six healthy recreational runners (height 1.72±0.09 m; mass 74±15 kg). A modified Cleveland Clinic marker set, including four-marker rigid-shell clusters on the shanks and thighs, comprising of 80 retroreflective spherical markers were placed on the subjects. Following a warm-up, subjects performed one 3s static-standing trial and ran for 2 mins at 3.35 m/s (7.5 mph) on a dual-belt treadmill instrumented with force platforms under each belt (TM-09-P; Bertec, Columbus, OH, USA) with a 30s trial recorded during this time. Kinetic data from the force platforms were amplified (16-bit AM6511; Bertec) and connected to the computer via an A-D board (16-bit NI-USB-6229; National Instruments, Austin, TX, USA). Using Cortex software (v2.0; Motion Analysis Corporation, Santa Rosa, CA, USA) the kinetic data were recorded at 1000 Hz and the three-dimensional motions of the markers were recorded at 200 Hz via eight cameras (4xEagle and 4xEagle-4; Motion Analysis Corporation). All procedures were approved by the institution's ethics review committee.
Data were analyzed using custom-written code in Matlab (v2012b; Mathworks, Natick, MA). Any small data gaps were linearly-joined, and kinematic and kinetic data were then smoothed using a fourth-order low-pass Butterworth filter at 5 Hz and 50 Hz cut-off frequencies, respectively. Heel-strike and toe-off events were determined using the force platform data rising above and falling below 30N, respectively. Data from heel-strike to subsequent heel-strike of the same leg (0% to 100% of stride time) were time normalized to 101 data points using a cubic spline interpolation. Using the thigh and shank clusters and four foot markers (heel lateral and posterior; first and fifth metatarsal), the three-dimensional joint coordinate systems (JCS) for the thigh, shank and foot were calculated.
The JCS were normalized to the anatomical standing position (i.e. 0° is angle during standing) and the knee flexion-extension (knee) and ankle dorsi-plantar flexion (ankle; -ve dorsiflexion; +ve plantarflexion) angles were calculated. Applying both 1D and 2D techniques, a 2-stage outlier detection method was applied to each subject separately as follows:

- Stage 1: at each time point the mean and SD across the strides was calculated, and any strides with a point exceeding limits of ±3.3SD (i.e. 99.9% area under a normal distribution) away from the mean were removed;
- Stage 2 using a moving window size of k=3:

I. Each cycle at the start (and end) was padded by k points using reflection of the first k (and last k) data points;
II. The strides were detrended by removing the mean at each time point to reduce variability between time points;
III. At each time point and incorporating the ±k points either side the moving window SD (mwSD) across the strides was calculated;
IV. Any stride that contained a data point exceeding ±2.58mwSD (i.e. 99.0%) from the detrended mean of zero were removed.

At each stage, the normal distribution of the data were verified using Shapiro-Wilks ($p>0.05$). The number of strides removed at each stage was described as means±SD. For each angle of each subject, the mean and SD at each data point were obtained and then a single mean calculated for the stride. The mean of these for the 6 subjects provided the group mean, which was compared between the raw, stage 1 and stage 2 data using a 1-way repeated measures ANOVA with LSD post-hoc analysis at an alpha level of 0.05.

**RESULTS:** From the 24 combinations of 6 subjects and 4 variables, Figure 1 illustrates one set of data. In this example, there are samples of both obvious outliers from incorrectly identified single-points and more subtle potential outliers as unrepresentative time-series (left). These outliers are removed after application of the 2-stage outlier detection method providing trials with all strides sharing a similar pattern with no obvious outliers (right).

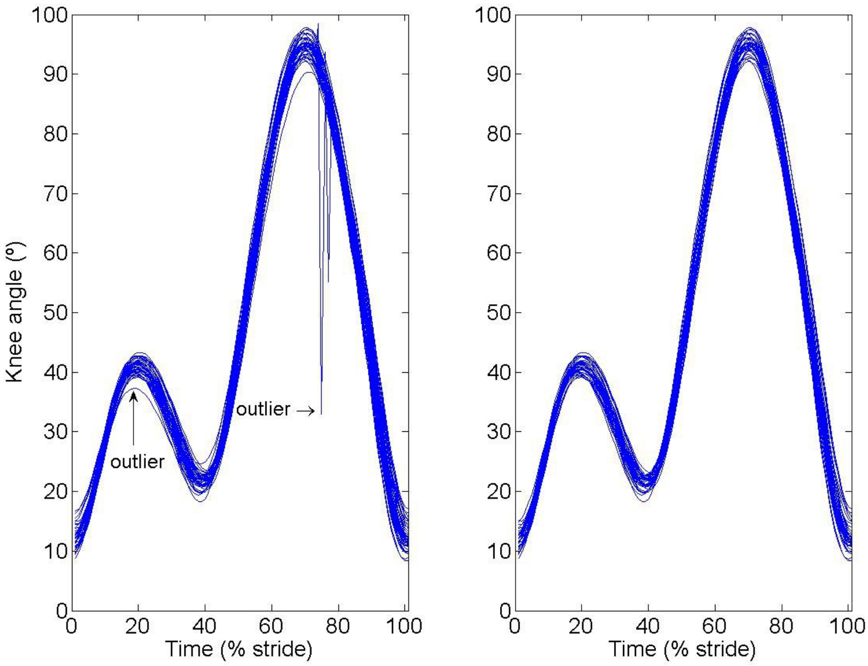


**Figure 1: Knee flexion-extension angle for 1 subject treadmill running at 3.35 m/s before (left; strides=44) and after (right; strides=37) a 2-stage outlier removal method. The data were the left leg for subject 3 from heel strike (0%) to heel strike (100%). The stance phase is approximately 0-40%. Some potential outliers on the left are indicated that were removed on the right.**

Both stages of the outlier detection method resulted in outliers being removed. There were 5.2±3 and 2.0±1.4 strides removed in stages 1 and 2, respectively. Primarily, the outlier removal did not alter the mean trace, although for the right ankle there was a significant but tiny change in the value (<0.1°). For all variables, the 2 stages combined resulted in a statistically significant reduction in variability (Raw to Stage 2; $p<0.05$) but the reduction between each stage was not always statistically significant (Table 1).

**Table 1: Number of strides and descriptive statistics (mean±SD) of the leg kinematics for six healthy subjects treadmill running at 3.35 m/s before (Raw) and after each stage of an outlier removal method (Stages 1 and 2).**

| Variable | Strides | | | Mean (°) | | | SD (°) | | |
|---|---|---|---|---|---|---|---|---|---|
| | Raw | Stage1 | Stage2 | Raw | Stage1 | Stage2 | Raw | Stage1 | Stage2 |
| Lank | 40.8±2.6 | 33.8±2.6 | 32.3±2.3 | 0.1 | 0.1 | 0.2 | $1.3^{c}$ | $1.2^{a}$ | 1.1 |
| Rank | 41.0±2.8 | 36.5±3.7 | 33.5±2.9 | $1.9^{c}$ | 1.9 | $2.0^{b}$ | $1.2^{c}$ | $1.1^{a}$ | $1.1^{b}$ |
| Lknee | 41.0±2.8 | 36.2±3.1 | 34.8±2.6 | 46.6 | 46.6 | 46.6 | $2.3^{c}$ | 2.2 | 2.2 |
| Rknee | 41.0±2.8 | 36.7±3.2 | 34.5±3.1 | 43.2 | 43.2 | 43.2 | $2.3^{c}$ | 2.1 | $2.0^{b}$ |
| Removed | | 5.2±3 | 2.0±1.4 | | | | | | |

Variables are: Left (L); Right (R); Ankle dorsi-plantar angle (ank); Knee flexion-extension angle (knee); strides removed at each stage (Removed). Mean and SD values are from the mean of the 6 subjects' mean values across the 101 data points of the strides. Significant differences ($p<0.05$) are indicated with: a (Raw v Stage 1); b (Stage 1 v 2); c (Raw v Stage 2). No statistical tests were applied to the number of strides.

**DISCUSSION:** The initial motivation for the development of the technique was to simplify the removal of errors from multiple strides of a single-subject. As seen in Figure 1, there are a few obvious errors indicated by sharp spikes, and these were detected in stage 1 of the method and the strides removed. A second need arose, whereby in the remaining strides, although following the general trend of the stride, there were more subtle differences such as mis-alignment of peak values, varying gradients in places, or parts of traces slightly offset from the group. These differences, particularly misaligned peak values and varying gradients, influence statistical descriptions over the second dimension of time. Hence, by stage 2 incorporating the variability across several points, strides with more subtle differing patterns to the group of trials were detected as outliers using the moving window SD component.

The primary assumptions in the method proposed are that the data possess a normal distribution and that there is similarity between the multiple strides. The first assumption is typically met by a sufficient number of trials, and verified with normality tests. The second assumption constrains this method to intra-subject outlier detection. However, if a task is novel and greater variability between strides is expected then this method may not be appropriate. Particularly with inter-subject analyses, the greater variability that exist would require that an alternative to this 2-stage method need proposing and evaluating.

There are several methodological choices that influence the detection of outliers. First, the method of outlier detection was based on SDs, and others may be advocated (e.g. IQR, MAD) although anecdotally we have found they provide comparable results. Second, and more importantly, the settings of the moving window size (k) and limits/area encompassed (e.g. 3.3SD, 99.9%; 2.58SD, 99.0%) have a much greater influence on the results. With k, as an odd value to prevent a phase-shift, the minimum of 3 was found effective. If the frequency content is low, such that the peaks and troughs flatten out more, then a higher value for k may be appropriate. With the limits, a higher threshold was applied at stage 1 to identify obvious discrete outliers, and to minimise removal of points at local minima or maxima that were not in alignment with each other across strides. In stage 2, as the strides should be more similar, a smaller threshold was warranted. Different k and limit settings, and even replacing SD with MAD, could be explored to reflect the data particularly in cases with fewer trials where the normality assumption is less well met, which may lead to a better removal of any outlier trials.

A primary benefit of the method it that it offers a simple way to remove outliers with no 'modification' techniques required (e.g. rectification, normalisation, transformation). This results in the remaining strides providing a more valid representation of the movement. Indeed, there is a possibility to use the method on an iterative basis to reduce the number of trials down further (e.g. to 5), whilst maintaining a normal distribution, and then to take the mean of these remaining trials as a better representation of the movement rather than being a distorted or 'mythical average' of the movement (Dufek *et al*., 1995).

**CONCLUSION:** This study describes a simple two-stage outlier detection method, which was found to be effective in removing outliers in intra-subject time-series data that are common in biomechanics.

**REFERENCES:**

Brownlees, C.T., & Gallo, G.M. (2006). Financial econometric analysis at ultra-high frequency: data handling concerns. *Computational Statistics and Data Analysis*, *51*, 2232-2245.

Dufek, J.S., Bates, B.T., James, C.J., & Stergiou, N. (1995). Interactive effects between group and single-subject response patterns. *Human Movement Science*, *14*, 301-323.

Helwig, N.E., Hong, S., Hsiao-Wecksler, E.T., & Polk J.D. (2011). Methods to temporally align gait cycle data. *Journal of Biomechanics*, *44*(3), 561-566.

Hodge, V.J., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*(2), 85-126.

Kaioti, V., & Caroni, C. (2004). Simple detection of outlying short time series. *Statistical Papers*, *45*, 267-278.

Mullineaux, D.R., Clayton, H.M., & Gnagey, L.M. (2004). Effects of offset-normalizing techniques on variability in motion analysis data. *Journal of Applied Biomechanics*, *20*, 177-184.