Physiol. Meas. 27 (2006) 413–423 doi:10.1088/0967-3334/27/4/007

A statistical test to determine the quality of accelerometer data

J E Slaven[1,4], M E Andrew[1], J M Violanti[2], C M Burchfiel[1] and B J Vila[3]

[1] Biostatistics and Epidemiology Branch, Health Effects Laboratory Division, National Institute of Occupational Safety and Health, Centers for Disease Control and Prevention, Morgantown, WV, USA
[2] School of Public Health and Health Professions, Department of Social and Preventive Medicine, State University of New York at Buffalo, NY, USA
[3] Criminal Justice Program and Sleep and Performance Research Center, Washington State University, Spokane, WA, USA

E-mail: cto8@cdc.gov

Received 11 January 2006, accepted for publication 20 February 2006
Published 14 March 2006
Online at stacks.iop.org/PM/27/413

Abstract
Accelerometer data quality can be inadequate due to data corruption or to non-compliance of the subject with regard to study protocols. We propose a simple statistical test to determine if accelerometer data are of good quality and can be used for analysis or if the data are of poor quality and should be discarded. We tested several data evaluation methods using a group of 105 subjects who wore Motionlogger actigraphs (Ambulatory Monitoring, Inc.) over a 15 day period to assess sleep quality in a study of health outcomes associated with stress among police officers. Using leave-one-out cross-validation and calibration-testing methods of discrimination statistics, error rates for the methods ranged from 0.0167 to 0.4046. We found that the best method was to use the overall average distance between consecutive time points and the overall average mean amplitude of consecutive time points. These values gave us a classification error rate of 0.0167. The average distance between points is a measure of smoothness in the data, and the average mean amplitude between points gave an average reading. Both of these values were then normed to determine a final statistic, K, which was then compared to a cut-off value, K_C, to determine data quality.

Keywords: actigraphy, quality control, sleep quality, statistical test

[4] Address for correspondence: MS 4050, 1095, Willowdale Rd, Morgantown, WV 26501, USA.

0967-3334/06/040413+11$30.00 © 2006 IOP Publishing Ltd Printed in the UK

1. Introduction

Using accelerometers to measure movement for the determination of sleep quality has been shown to correspond very well with polysomnography (Cole *et al* 1992, Jean-Louis *et al* 1996). However, data can be corrupted by the following: a malfunction of the hardware inside the accelerometer, an incomplete transfer of data from the accelerometer to a computer, and wearer non-compliance with the research protocol caused by taking off the accelerometer more often than instructed.

Sleep measurements are determined by applying researcher-selected algorithms to activity signals that are generated in the accelerometer by user movement. These signals are downloaded into the computer software and displayed as reading channels. The three most common reading channels are the zero-crossing mode (ZCM), time-above-threshold mode (TAT) and proportional integrating measure mode (PIM). Each of these modes measures movement, either as frequency of motion, as time spent in motion or as vigor of motion, respectively.

The Motionlogger actiwatch also has a life channel. This is an extremely sensitive, wide-band channel. It captures micro-vibrations originating from the central tremor of the body, from pulse and from respiration as well as overall movement. Due to the amount of information it receives it is not useful in determining sleep times, but is useful for quality control in showing when the actiwatch was on the subject during the study. Life channel values of very low to zero amplitudes indicate that the actiwatch is not being worn.

In a sample of 105 participants from a study of police officer sleep quality, we noticed that the life channel had gaps in the data. The question arose as to how often these gaps occur and how long the actiwatches can be off a subject's wrist before significantly affecting sleep statistics. To our knowledge, few studies have reported quality control of accelerometer data, and no actual tests have been discussed (Van Coevering *et al* 2005). In order to assess the quality of accelerometer data, the authors in the cited study used quality control functions that came with the statistical analysis software they used. However, not all such programs come with quality control features. As the use of accelerometers is becoming more prevalent in sleep quality and physical activity studies, it was important to identify a reliable and robust statistical test for determining if the quality of accelerometer data is sufficient for use in statistical analyses. Visual inspection of the data can also be done to determine quality, but large sample sizes and data with many days of records can make this approach prohibitive. A faster and more automated approach that reduces the analysis time would be useful.

2. Methods

A study of health outcomes associated with stress among police officers includes analysis of one of the research questions under study: how stress affects sleep quality? To help ascertain this, officers were asked to wear an accelerometer in order to record their movement, which allows us to determine the quantity and quality of their sleep. The Motionlogger actigraph was selected as an accelerometer. It records data using the channels outlined above, allowing sleep quality statistics to be derived. The accelerometer also contained a time piece, allowing it to be used as a wristwatch as well. This area of placement also enables the accelerometer to detect acceleration along all three dimensions of movement. This investigation uses data from the first 105 of an anticipated 700 officers from the Buffalo Police Department. Data from the remaining 595 officers have not yet been collected. A general description of the study's design, methods and participant characteristics has been reported (Violanti *et al* 2006). Initially, the Motionlogger actigraphs were to be worn for 15 days without taking them off at

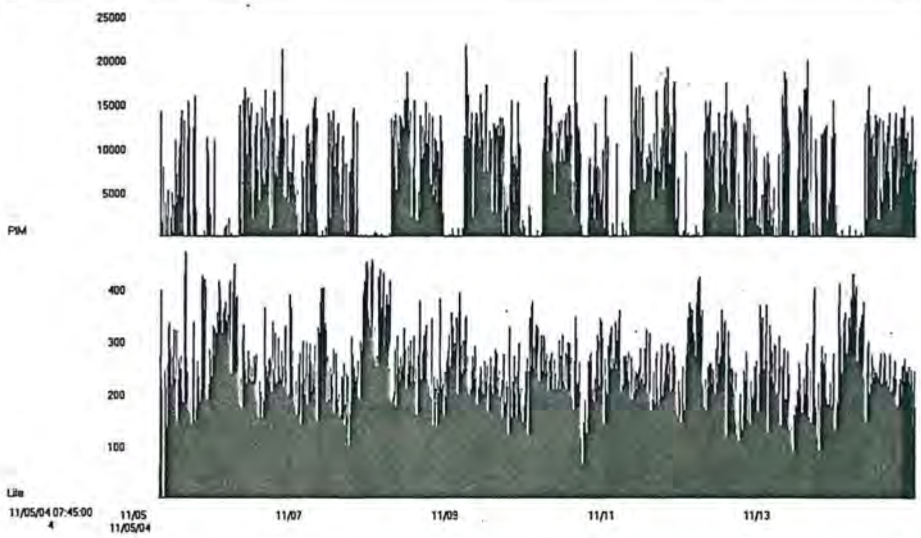

Figure 1. Good quality life channel reading. Note the periodicity of the PIM channel and no-low readings of the life channel, indicating subject compliance and no-data corruption. The *X*-axis is time/date; the *Y*-axis is PIM channel and life channel readings in volts.

any time. It was found that the models were not completely water resistant, and the protocol was changed to allow the users to take them off for short periods of time in order to protect them from water damage (e.g., while bathing or swimming). All phases, testing and reports of the study were approved by the State University of New York at Buffalo Internal Review Board and the National Institute for Occupational Safety and Health Human Subjects Review Board.

The data from the accelerometers were transferred into computer files using Action4 software (Ambulatory Monitoring, Inc.). Sleep was scored using the primary integration mode channel. PIM has previously been shown to work at least as well as the zero-crossing mode and the time-above-threshold methods in scoring sleep (Gerardin *et al* 2001). Each of the original Action4 files was visually inspected and given a quality rating of good or poor based on the length and number of low readings in the life channel and Action4's generated sleep statistics. The files were then exported into Excel and finally transformed into SAS data files. Data analysis was performed using SAS v. 9.1 (SAS Institute, Cary, NC).

Several methods were assessed to find a usable quality control statistic for the accelerometer data. Initially, autocorrelation, periodicity and spectral analysis measures were examined. These are useful for determining sleep quality and standard statistics, but none of these measures is capable of assessing data quality. Poor quality data still show much the same correlation, cosinor waves and spectral analysis as good quality data, due to the inherent periodicity of sleep–wake cycles.

One of the two main features of corrupted and non-compliant data is a large decrease in the life channel resulting in very low to zero readings due to the watch being off the participant, and recording only readings from nearby vibrations. The other feature is a flat line on the life channel due to data corruption, which can occur in the watch itself due to mechanical error or from a transfer error when the data are downloaded from the accelerometer into a computer. Figure 1 shows an example of good quality life and PIM channel readings.

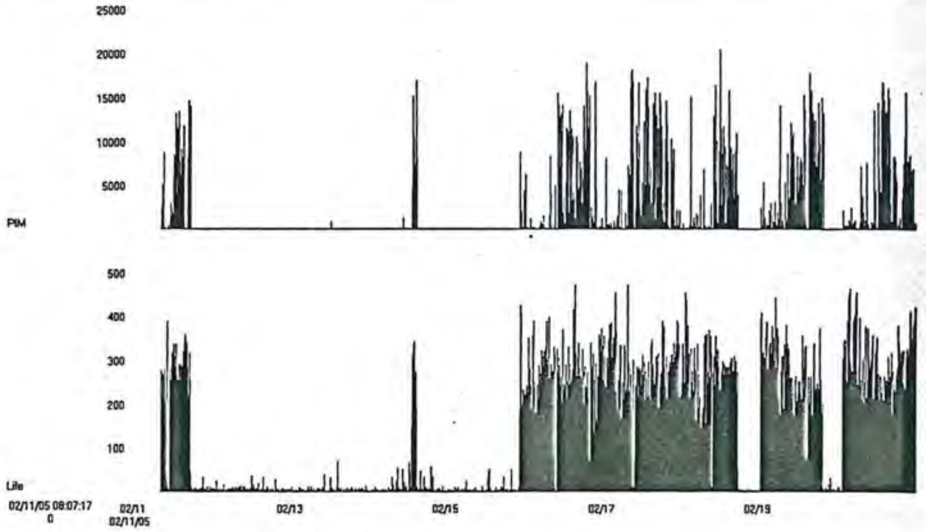

Figure 2. Poor quality life channel reading due to non-compliance. Note the several days where the life channel reading is at or near zero, indicating that the accelerometer has not been worn. The *X*-axis is time/date; the *Y*-axis is PIM channel and life channel readings in volts.

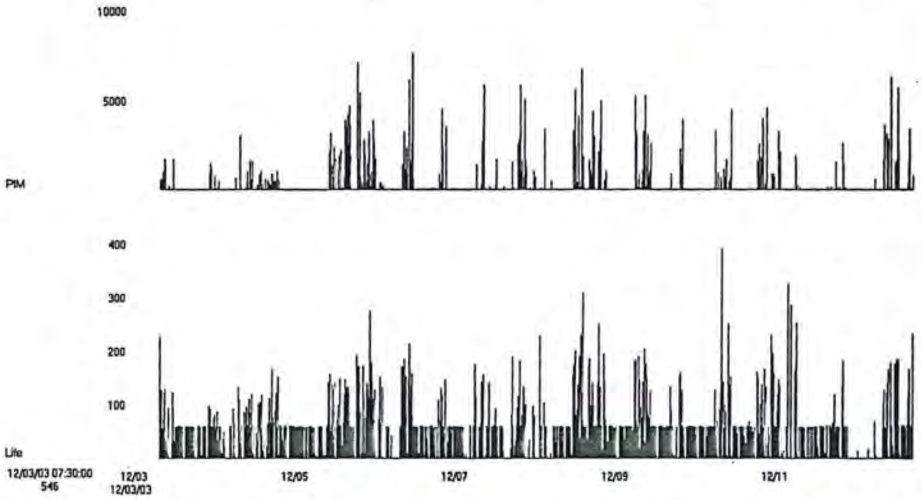

Figure 3. Poor quality life channel reading due to data corruption. Note the long time periods of straight readings on the life channel, indicating that the data have been cut off at some point. The *X*-axis is time/date; the *Y*-axis is PIM channel and life channel readings in volts.

Figure 2 illustrates what happens when non-compliance occurs. Figure 3 shows data that have been corrupted inside the accelerometer or during downloading into a computer. Figure 4 shows a combination of non-compliance and data corruption.

In order to test for large decreases in life channel readings, the mean life channel reading was taken. If there are many readings at or near zero, it is an indication that the officer was

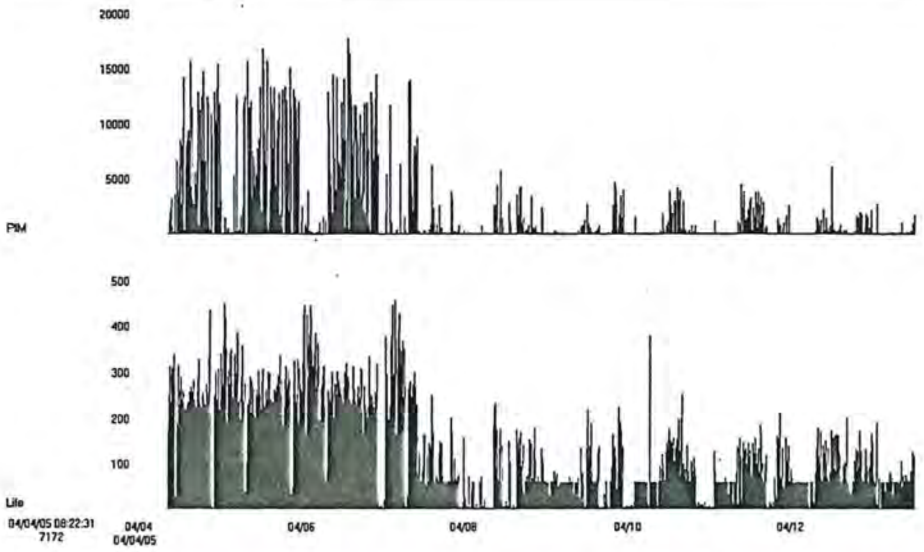

Figure 4. Poor quality life channel reading. There are a few portions with a near-zero reading and several linear portions indicating that the data have been truncated. The *X*-axis is time/date; the *Y*-axis is PIM channel and life channel readings in volts.

not wearing the accelerometer and would result in a low mean. Very few zero or near zero readings show compliance and result in a higher mean. Another form of the mean was also determined by calculating the average amplitudes between consecutive time points and then determining the mean of all such averages.

To find if there is a significantly large area where the life channel readings are flat, which would indicate data corruption, the variance, skewness and distance between consecutive time points were examined. A low variance means little change in readings, indicating a flat line. Skewness gives a reading of how much the life channel readings are above or below the average. If these data remain at a very low reading for a long period of time, it is due to the readings being cut off at the same point, indicated by a positive skewness. Averaging the distance between amplitudes of consecutive time points measures data smoothness. For relatively flat readings, the distances between each consecutive time point would tend to be small, quantified by small overall average of the distances. This is illustrated in figure 2, where there is little or no change between consecutive data points, indicating very little distance in the display of the data. Figure 1 shows that good quality data have larger distances between consecutive time points.

In order to identify the optimal technique for assessing the quality of life channel data, we calculated the mean, variance, skewness, overall mean amplitude of consecutive time points and the overall mean distance of consecutive time points, paired them together, and then generated graphs. Although these are standard statistical calculations, the formulas for the sample statistics are given below for reference.

$$\text{Mean} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$\text{Variance} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2,$$

$$\text{Skewness} = \frac{1}{(n-1)s^3} \sum_{i=1}^{n} (x_i - \bar{x})^3,$$

where s is the sample standard deviation.

$$\text{Average amplitude of consecutive time points} = \frac{1}{n-1} \sum_{i=1}^{n-1} \left(\frac{x_i + x_{i+1}}{2} \right).$$

$$\text{Average distance between consecutive time points} = \frac{1}{n-1} \sum_{i=1}^{n-1} \left(\frac{x_i - x_{i+1}}{2} \right).$$

The pairs were mean–variance, mean–skewness, one triple with mean–variance–skewness and average distance–average amplitude between time periods. Visual inspection of the graphs allowed us to compare how well the good data and the poor data were separated by a method.

The paired variables were also analyzed using discriminant analysis to see how well they were separated using a discrimination function by giving an error rate of how many were misclassified. The discrimination function provides a criterion for group assignment between the two categories of good and poor quality. A generalized squared distance function, $D_i^2(x)s = (x - \bar{x}_j)' \operatorname{cov}_j^{-1}(x - \bar{x}_j)$, is used to find the posterior probability of group membership, given by $\Pr(j|x) = \exp\bigl(-0.5D_j^2(x)\bigr) \big/ \sum_k \exp(-0.5D_j^2(x)\bigr)$. Error rates are then determined by finding the number of data points classified differently from their original group.

Two methods were used in this procedure. The first was to use the leave-one-out cross-validation method (Hastie *et al* 2001). A discrimination function was calculated by taking all but one of the data points, and then the remaining data points were classified using that function. This was then repeated for all of the data points in the set. The second method was the calibration-testing method (Hastie *et al* 2001). As recommended, a random 20% of the data points were used as a calibration set to calculate the discriminatory function. The other 80% were then analyzed using that function. A random 25% of the data points were later used as a calibration set to test for sensitivity. The results are similar to the original test using 20%.

After identifying the variable pair with the smallest error rate, we tested again, but this time only using the data obtained during a subject's sleep cycle. This removed all of the zero life channel readings that are present during wakefulness. Both methods of discrimination analysis were used on this data set to see if using only the data obtained while the subject was asleep was a better way for determining data quality.

For each of these five pairs of variables, a final quality statistic, K, was then calculated by taking the norm of the variables used. The norm in this case is the Euclidean distance between the origin and the ordered pair formed by the variables (equation (1)), where x and y each represent one of the variables in the pair. Which variable is x and which is y does not matter. The ordered triple obtained by using the mean, variance and skewness also has a z component.

$$K = (x^2 + y^2)^{1/2}. \tag{1}$$

Using the K-statistic will separate the data into clusters of good and poor quality. By graphing the data points, it is seen that good quality data points cluster toward the right of the graph. Poor quality data points will be to the left and more dispersed around the graph. There will be some intersection of the two quality types, as demonstrated by the error rate in the discrimination analysis.

A cut-off value, K_C, is then identified by visual inspection of the graph. Due to the different placement of the good quality data-point cluster and the poor quality data-point cluster, there will be a noticeable separation between the two groups. This separation will be

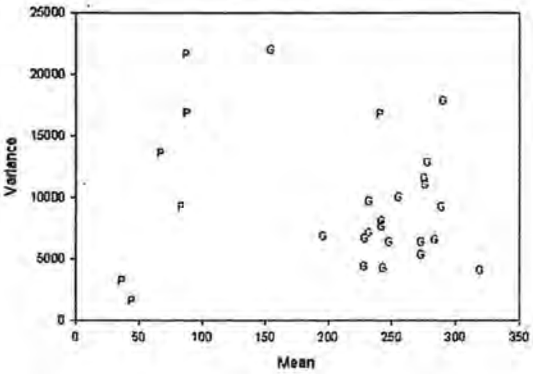

Figure 5. Mean and variance of life channel readings. G = good quality data points, P = poor quality data points. Units are in volts.

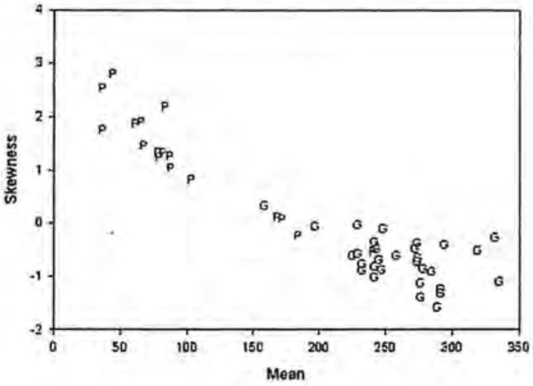

Figure 6. Mean and skewness of life channel readings. G = good quality data points, P = poor quality data points. Units are in volts.

more noticeable depending on which approach is used. K_C is then calculated by taking the mean of the poor quality data point and the good quality data point that are closest together. Figure 8 gives the best example of how the data separates and how large the gap between good and poor quality can be. If $K > K_C$, then the quality of the data is good and should be kept. If $K < K_C$, then the quality of the data is poor and should be analyzed to determine if anything can be done to improve its quality or whether it should be discarded.

3. Results

Error rates for the leave-one-out cross-validation method ranged from 0.0333 to 0.0778 while those for the calibration-testing method ranged from 0.0167 to 0.4046 (table 1).

The method with the least error for categorizing data quality is the overall average amplitude between consecutive time points and the overall average distance between consecutive time points. It has the lowest error rate for both methods of discrimination. There is a very large difference in the error rates for the calibration-testing discrimination analysis between this method and the others, with an error rate of less than 2%. Figures 5–9

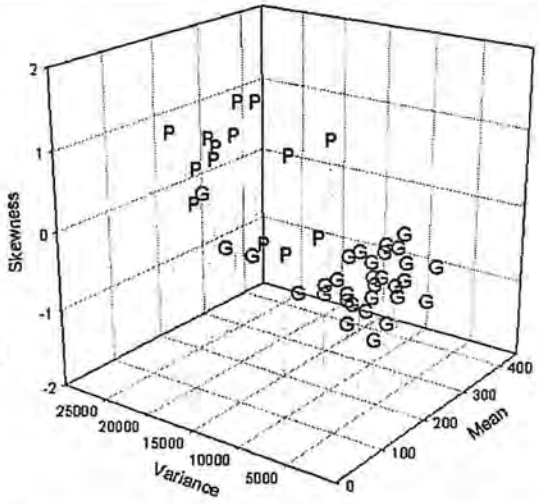

Figure 7. Mean, variance and skewness of life channel readings. G = good quality data points, P = poor quality data points. Units are in volts.

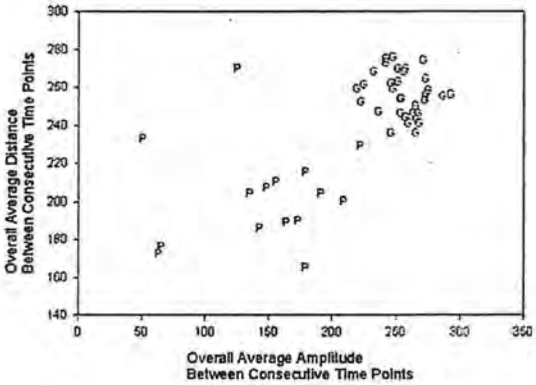

Figure 8. Average of the distance between consecutive time points and the average of means between consecutive time points of life channel readings. G = good quality data points, P = poor quality data points. Units are in volts.

graphically show that this method has the best separation between good and poor quality data over the other methods. The labels in the graph are the quality designations given from the initial visual inspection. P is for poor quality data and G is for good quality data. Although there is some overlap between the two quality type clusters, there is very little in this method (figure 8). This method works well with the signal structure of the actigraph by acting as a smoothing function. It determines differences between consecutive points, which allows an analysis of epoch by epoch change, instead of giving one parameter that tries to represent the entire fluctuating data set.

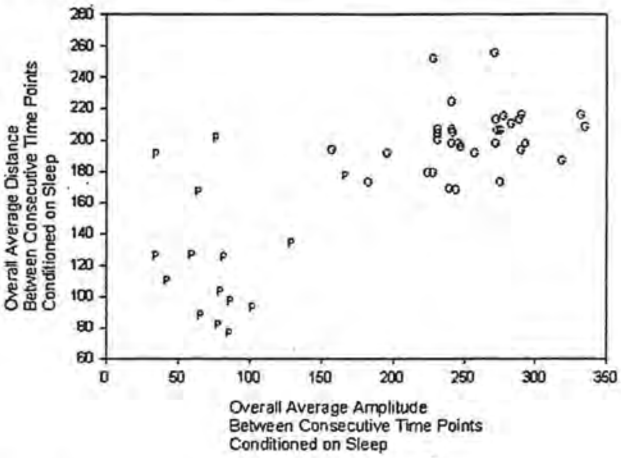

Figure 9. Average of the distance between consecutive time points and the average of means between consecutive time points conditioned on sleep of life channel readings. G = good quality data points, P = poor quality data points. Units are in volts.

Table 1. Comparison of error rates for leave-one-out cross-validation and calibration-testing discrimination.

Approach	Leave-one-out Error rate	Calibration-testing Error rate
Mean and variance	0.0778	0.12
Mean and skewness	0.0617	0.4046
Mean, variance and skewness	0.0778	0.1769
Overall average distance and amplitude	0.0333	0.0167
Overall average distance and amplitude conditioned on sleep	0.0333	0.0833

The K-statistic for this method is given by equation (2), where the overall average distance between consecutive time points is the x-coordinate from equation (1) and the overall average amplitude is the y-coordinate from equation (1).

$$K = ((\text{overall-avg-dist})^2 + (\text{overall-avg-amplitude})^2)^{1/2}. \qquad (2)$$

The largest norm value for the poor quality data cluster was 319 and the smallest norm value for the good quality data cluster was 341, giving a cut-off value of $K_C = 330$.

4. Discussion

We can categorize accelerometer data into quality groups with an extremely low error rate by using the pair of the overall average distance and average amplitude between time periods in conjunction with its K-statistic. This technique provides a standardized tool to help researchers avoid using poor quality actigraph data, whether they arise from file corruption or subject non-compliance, in order to prevent inaccurate sleep quality assessments.

From a practical standpoint, this technique is also efficient because it allows the researcher o accurately determine K_C by using only a small percentage of the data set. Once this is done, it is relatively easy to categorize remaining data points. This may be important because it may be possible to salvage blocks of good data. Note that several studies have used anywhere from 2 to 39 days for data collection (Acebo *et al* 1999, Park *et al* 2000, Lo *et al* 2002, Blood *et al* 1997). Although studies involving adults are needed, for children and adolescents it is indicated that 1 or 2 days of data are unreliable and that 5 days or more are adequate (Acebo *et al* 1999, Sadeh and Acebo 2002). When subjects are on shift work, it may also be desirable to collect high quality data over the course of a full-shift cycle.

Using this technique will also allow data analysts to be more efficient in reviewing data. Only the data that are categorized into the poor quality group would need to be examined to determine the cause of poor quality, rather than looking at each subject's data. It may also aid in determining what caused the poor quality. The *K*-statistic data cluster in such a way that the points in the lower left-hand corner of the graph typically come from data files that have been corrupted, while the data in the upper left-hand corner of the graph come from subjects who did not properly follow protocols. This appears to be due to the low variance of corrupt data, since those data not only have low readings, but also readings that do not change from epoch to epoch. In our experience, poor quality data points near the cluster of good data points typically have more low life channel readings than normal, and they also usually had a large section at the end of the data file with low life channel readings. This occurred when subjects finished the study and removed the accelerometer, but it was not turned off. There were not enough subjects in our analyses to significantly show if different types of poor data will cluster separately. Although the area separating the poor quality data and the good quality data should be noticeable, there may be data points that fall very close to the cut-off value that would need to be examined.

The entire subject group is composed of police officers. The statistics for sleep quantity and quality may be different than other professions or in more heterogeneous groups. However, this method of determining overall data quality is valid for any subject group as it gives information on data quality detection in general, not the quality of sleep.

This method of quality control may also be used for other purposes where the output is a continuous stream of data, such as production equipment and medical monitors.

In conclusion, using the *K*-statistic will categorize the quality of accelerometer data as either good or poor with a very low error rate. Use of this statistic will save analysts time and aid in the decision of which segments of data should be used.

Acknowledgments

This work was supported by the National Institute for Occupational Safety and Health (NIOSH), contract no 200-2003-01580. The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the National Institute for Occupational Safety and Health.

References

Acebo C, Sadeh A, Seifer R, Tzischinsky O, Wolfson A, Hafer A and Carskadon M 1999 Estimating sleep patterns with activity monitoring in children and adolescents: how many nights are necessary for reliable measures? *Sleep* **22** 95–103

Blood M, Sack R, Percy D and Pen J 1997 A comparison of sleep detection by wrist actigraphy, behavioral response and polysomnography *Sleep* **20** 388–95

Cole R, Kripke D, Gruen W, Mullaney D and Gillin J 1992 Automatic sleep/wake identification from wrist activity *Sleep* **15** 461–9

Gerardin J, Kripke D, Mason W, Elliott J and Youngstedt S 2001 Sleep estimation from wrist movement quantified by different actigraphic modalities *J. Neurosci. Methods* **105** 185–91

Hastie T, Tibshirani R and Friedman J 2001 *The Elements of Statistical Learning, Data Mining, Inference and Prediction* (New York: Springer)

Jean-Louis G, von Gizycki H, Zizi F, Fookson J, Spielman A, Nunes J, Fullilove R and Taub H 1996 Determination of sleep and wakefulness with the actigraph data analysis (ADAS) *Sleep* **19** 739–43

Lo C-C, Amaral L, Havlin S, Ivanov P, Penzel T, Peter J-H and Stanley H 2002 Dynamics of sleep-wake transitions during sleep *Europhys. Lett.* **57** 625–31

Park Y, Matsumoto K, Seo Y, Cho Y and Noh T 2000 Sleep-wake behavior of shift workers using wrist actigraph *Psychiatry Clin. Neurosci.* **54** 359–60

Sadeh A and Acebo C 2002 The role of actigraphy in sleep medicine *Sleep Med. Rev.* **6** 113–24

Van Coevering P, Harnack L, Schmitz K, Fulton J, Galuska D and Gao S 2005 Feasibility of using accelerometers to measure physical activity in young adolescents *Med. Sci. Sports Exerc.* **37** 867–71

Violanti J *et al* 2006 The Buffalo Cardio-Metabolic Occupational Police Stress (BCOPS) pilot study: methods and participant characteristics *Ann. Epidemiol.* **16** 148–56