Stepping Up: Improving Step Count ML Algorithms

Brendan Callender, bscallen@calpoly.edu
Jadyn Ellis, jellis13@calpoly.edu
Martin Hsu, mshsu@calpoly.edu
Kirina Sirohi, kasirohi@calpoly.edu
Instructors: Dr. Hunter Glanz, Dr. Jonathan Ventura
Clients: Dr. Sarah Keadle, Paige Dolan
DATA 452, Spring 2024

I. Introduction

Step counting is essential for measuring physical health and human activity, as higher daily step counts are associated with a lower risk of heart disease, cancer, and death[1]. Wrist-worn accelerometers provide a low-burden method for individuals to measure daily steps; however, certain activity classifications that involve restricted or increased wrist movement can amplify algorithm step prediction errors. Such classifications include “modified walking”, which encompasses walking with a load or ascending and descending stairs, where prediction accuracy is often depleted. Moreover, signals associated with certain postures tend to behave different from others, so peak detectors may fail in step prediction in instances of unfamiliar signals. Errors in step prediction by existing algorithms can adversely impact the measurement of daily physical activity, ultimately misinforming individuals about their daily physical activity levels. We aim to improve the performance of an existing step-counting algorithm by the UK Biobank, specifically aiming to enhance step prediction accuracy during modified walking periods and developing an optimal set of postures for multiple trained peak detectors.

II. Background

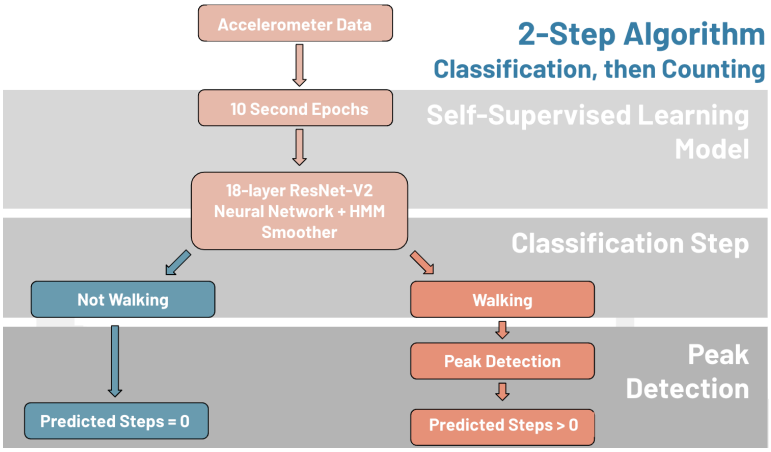
The complete data we were given includes two parts: accelerometer features and manually labeled ground truth. The Activities Completed over Time (ACT24) is a tool that captures accelerometer data in 3-D space at 80 readings per second (80Hz). Throughout a 7-day period, 53 participants were asked to wear an activPal device and complete three ACT24 recalls. Of those 53 participants, 24 consented to two 3-hour recorded sessions which provided a crucial ground truth reference to serve as a reliable training label for supervised machine learning. The ground truth was recorded by Dr. Sarah Keadle’s team at the Cal Poly Kinesiology Department through manual labeling of the recorded sessions to classify activity and posture levels as well as count the number of steps taken.

[1] https://www.nih.gov/news-events/nih-research-matters/number-steps-day-more-important-step-intensity

Dr. Keadle selected a step-counting algorithm from the UK Biobank as a benchmark machine learning algorithm for our team to study and evaluate. The UK Biobank is one of the largest cohort studies globally, comprising health-related data from over 500,000 participants aged between 40 and 69 years. One of the key components of this UK Biobank data is the collection of accelerometer data from wrist-worn devices, similar to the activPAL device worn in the ACT24 study. By employing machine learning techniques and leveraging a ground truth annotated dataset, researchers were able to develop a more accurate step-counting algorithm. We were then able to employ this algorithm on our ACT24 data to serve as a baseline algorithm that we are hoping to improve.

The UK Biobank algorithm is structured as a 2-step algorithm[2], involving a classification step and a peak detection step. First, the algorithm classifies 10-second intervals of data as either periods of walking or not walking, where periods of walking are defined as 10-second intervals with 4 or more steps[3]. After classification, the algorithm then applies a peak detection algorithm to count steps. The peak detection algorithm is only applied to 10-second intervals predicted as walking, and non-walking intervals are automatically assigned 0 steps.

Figure 1. UK Biobank Algorithm Structure



III. Data Description

The data consists of accelerometer and ground truth data. The accelerometer data, recorded by an activPAL device, includes the time, x, y, and z coordinates, capturing 3-axis movement at 80 readings per second. The ground truth data is split into two files. The first file, "behavior/posture", contains information about ID, Session, Time, Activity, and Posture, providing details on the participants' activities and postures. The second file, "steps", includes

[2] https://github.com/OxWearables/stepcount
[3] https://oxwearables.github.io/ssl-wearables/

information about ID, Session, Time, and Steps, documenting the number of steps taken per second.

IV. Methods

The first part of the project involved data cleaning and exploration. Firstly, we prepared a version of the ACT24 dataset that matched the specifications and basis needed to run the UK Biobank algorithm. After testing the original algorithm loaded from the UK Biobank repository on the ACT24 data, we developed timeline plots and calculated error metrics that allowed us to both visually and analytically inspect the algorithm for areas of weakness. We particularly focused on systematic underestimation of periods of modified walking. When run on ACT24 data, the UK Biobank model was off by 2.02 steps per minute for the entire dataset. However when looking at periods of modified walking, the error increases to 22.9 steps per minute (*Table 1*).

Table 1. UK Biobank Algorithm Absolute Error per Minute on ACT24 Data

Overall	During Periods of Modified Walking
2.02 Steps	22.90 Steps

After identifying these areas of weakness, we directed the focus of our efforts this quarter toward editing and improving the UK Biobank algorithm. Specifically, we decided to investigate whether using more granular posture classification would improve the model. Additionally, we aimed to determine the optimal level of granularity to minimize classification errors without risking overfitting. We accomplished this by adapting the "classification, then peak detection" approach taken by the UK Biobank algorithm to handle classification of three different levels of granularity of walking movement – broad, granular, and condensed granular (*Table 2)*. To put each model on the same basis of comparison, we trained each model on the ACT24 data and performed both leave-one-individual-out cross-validation (LOOCV) and testing using the external OxWalk dataset from the UK Biobank.

Table 2. Breakdown of algorithms

Method	Training Data	Classification	Peak Detection
UK Biobank	OxWalk	Walk vs Not Walk ≥ 4 steps	Peak detection on periods of walking
Broad Classification	ACT24	Walk vs Not Walk ≥ 4 steps	Peak detection on periods of walking
All Posture Classification	ACT24	All Postures	Tuned peak detection on every posture

Condensed Posture Classification	ACT24	Condensed Set of Postures	Tuned peak detection on every posture

First, we defined the three levels of classification for the modified "classification, then peak detection" model – broad, granular, and condensed granular. The broad classification approach matched the original UK Biobank method of classifying 10-second epochs of accelerometer signal data into simple walking versus not walking categories. The granular classification approach classified epochs into all possible posture categories available in the ACT24 ground truth. Finally, the condensed granular approach classified the epochs into fewer categories than the granular approach, but more categories than the broad approach, representing a middle ground between the two. We created categories in the condensed approach by deriving reduced or aggregated categories from the granular categories, as seen in *Table 3*. The categories were grouped together into broader categories based on whether or not they were walking, modified walking, or not walking. Furthermore, they were condensed via a visual inspection for similarities in signal seasonality and magnitude, as well as similarities in peak distance and prominence metrics measured by fine-tuning a peak detection model on each granular category.

Table 3. Granular Posture to Condensed Posture Mapping

Condensed Posture	Granular Posture Equivalents/Definition
Walk	Walk
No Movement	Stand, Sit, Stretch, Kneel/Squat, Lying
Stand and Move	Stand and Move
Ascend Stairs	Ascend Stairs
Modified Walking	Descend Stairs, Walk with Load
Bike	Bike
Muscle Strengthening	Muscle Strengthening
Other Sport Movement	Other Sport Movement
Running	Running

After defining the different levels of classification, we modified the original UK Biobank algorithm to handle these classification levels. The UK Biobank algorithm consists of a neural network for classification with a hidden markov model smoothing process, followed by a single peak-detection algorithm that counts the steps for epochs classified as walking. We first modified

the neural network specification to classify more posture categories, then modified the peak detection process to include one peak detector fine tuned on each category.

Next, we obtained cross-validation metrics using a leave-one-individual-out deterministic method of cross-validation with ACT24 training data. This was performed as a first step to not only approximate overall test error on the ACT24 data, but also for each posture, which could only be done with the ACT24 training data. Each valid individual's 1-2 observed activity periods were treated as a single fold in the cross-validation process. Each fold would be treated as the holdout test set once, while the remaining folds would be treated as the training set, resulting in 19 cross-validation models trained and then tested total. The resulting step counts were aggregated into a single dataset, with which we evaluated estimated test error.

Finally, we trained a model at each classification level on the entire ACT24 dataset and tested each model on the OxWalk data. Because the OxWalk data does not include posture and activity classifications in its ground truth, we were only able to evaluate error on the overall count of steps, without finding the error by posture. However, this allowed us to validate our algorithm on new data entirely and confirm the error estimated by the cross-validation process.

For cross-validation on ACT24 data and testing on OxWalk data, we produced error metrics and plots which aided to compare results across all three classification levels. The results were compared on the basis of overall step counts as well as step counts broken out by posture or individual observation.

V. Results

Our study evaluated the performance of modified step counting algorithms through rigorous testing methods. We aimed to assess the effectiveness of our algorithms by employing leave-one-out cross-validation (LOOCV) on the ACT24 training data and utilizing the external OxWalk data as a testing set. Although using OxWalk as a testing set allowed us to train the model using all labeled postures from the ACT24 data, we encountered limitations as OxWalk has no labeled postures, preventing us from evaluating the models with respect to posture.

Similar to the structure of the UK Biobank algorithm, our three modified algorithms perform a classification step followed up by a peak detection step. *Table 3* contains the overall classification accuracy on the ACT24 data using LOOCV. We observe that the classification accuracy decreases as the number categories in the classification step increases. However, the decrease in classification accuracy from the Walk/Not Walk model is much smaller for the Condensed Posture model compared to the All Posture model. This supports our hypothesis of there being an effective middle ground between predicting the broad categories of walking vs not

walking and the granular approach of using all postures. Complete confusion matrices for each of the models can be found in the *Appendix*.

Table 4. LOOCV Classification Step Accuracy

Walk / Not Walk	Condensed Postures	All Postures
86.3%	82.6%	66.6%

LOOCV was also used to evaluate the success of the different models on the ACT24 for periods of modified walking and across the entire dataset. *Table 4* shows the absolute error per minute for each of the models overall and for periods of modified walking. Overall, the condensed postures model performed the best with an error of 2.58 steps per minute across the entire dataset followed by the all posture model with an error 3.21 steps per minute. Similar to the UK Biobank, each of the models had substantial increases in error for periods of modified walking with no approach outperforming the UK Biobank for periods of modified walking (22.90 steps/min).

Table 5. LOOCV Absolute Error per Minute on ACT24 Data

Category	Walk / Not Walk	Condensed Postures	All Postures
Overall	3.90 Steps	2.58 Steps	3.21 Steps
Modified Walking	25.46 Steps	25.85 Steps	24.71 Steps

Using OxWalk data as a test set allowed us to train each model on the entire ACT24 dataset to maximize the amount of training data with labeled postures. The OxWalk wrist-worn 100Hz data[4] only contains information about thirty-nine sessions, averaging about one hour in length, with no labeled postures. After running each fully trained model on the test set, we calculated the absolute error per minute across the entire test set. From *Table 6*, the condensed posture model and all posture model outperformed the UK Biobank with lower error values. For each of the 4 models, we observe a much lower error rate which is perhaps due to the OxWalk data being much less diverse with respect to the amount of postures performed in the data.

Table 6. OxWalk Test Set Absolute Error per Minute

UK Biobank	Walk / Not Walk	Condensed Postures	All Postures
0.79 Steps	3.81 Steps	0.40 Steps	0.23 Steps

[4] https://ora.ox.ac.uk/objects/uuid:19d3cb34-e2b3-4177-91b6-1bad0e0163e7

VI. Discussion

Overall when using the ACT24 data to train models, we observed that classifying more categories tends to improve step counting outcomes. This is evident in the lower overall rates when using LOOCV to test the models on the ACT24 data for the all posture and condensed posture models. Additionally, the more granular models had a significantly lower error rate when run on the OxWalk test set compared to walk/not walk model (See *Results*).

Our motivation for categorizing postures more extensively was primarily to enhance the peak detection step of our algorithm. Each potential predicted posture is associated with a specific peak detector optimized on the different signals within these predicted postures (*Figure 2*). Consequently, the expanded posture sets – all postures and condensed postures – tend to be more successful in predicting steps. While the classification error increases with more categories, these models yield more accurate step counts due to the improved accuracy of the peak detection process.

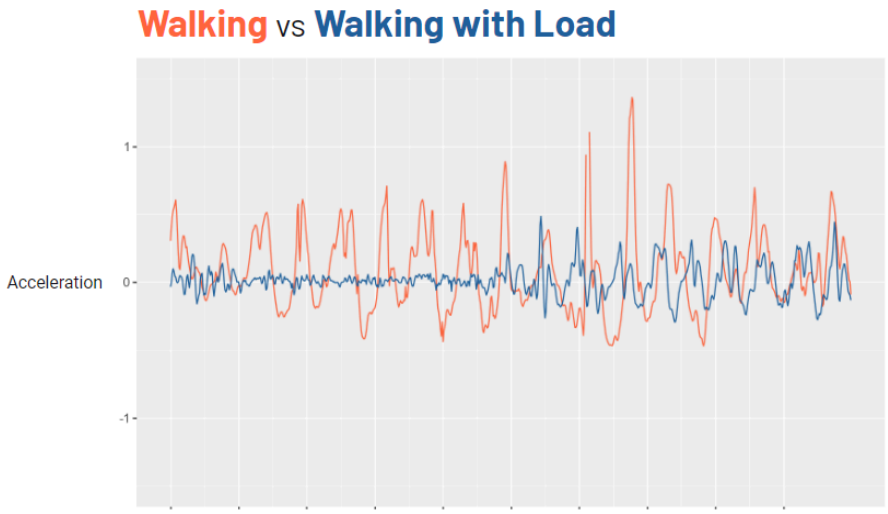


Figure 2. Acceleration signal comparison for walking and walking with a load

Additionally, breaking up the postures means the walking category no longer contains a variety of signal types. Since the Walk/Not walk model was the exact same as the UK Biobank model, except for being trained on ACT24 data, we believe this approach requires significantly more data for the classification algorithm to be able to correctly aggregate unique signals. By breaking up these different signals into more easily recognizable groups, we were able to achieve more accurate step counts with less training data.

In our previous work, our team identified a weakness in the UK Biobanks algorithm's ability to predict steps during modified walking periods (*See Figure 3)*. Despite our efforts to hone in on

modified walking step prediction, we were unable to significantly improve step prediction for modified walking periods. The absolute error per minute during modified walking periods was substantially higher for all three of our algorithms compared to the overall error rate (*See Table 4 in Results)*.

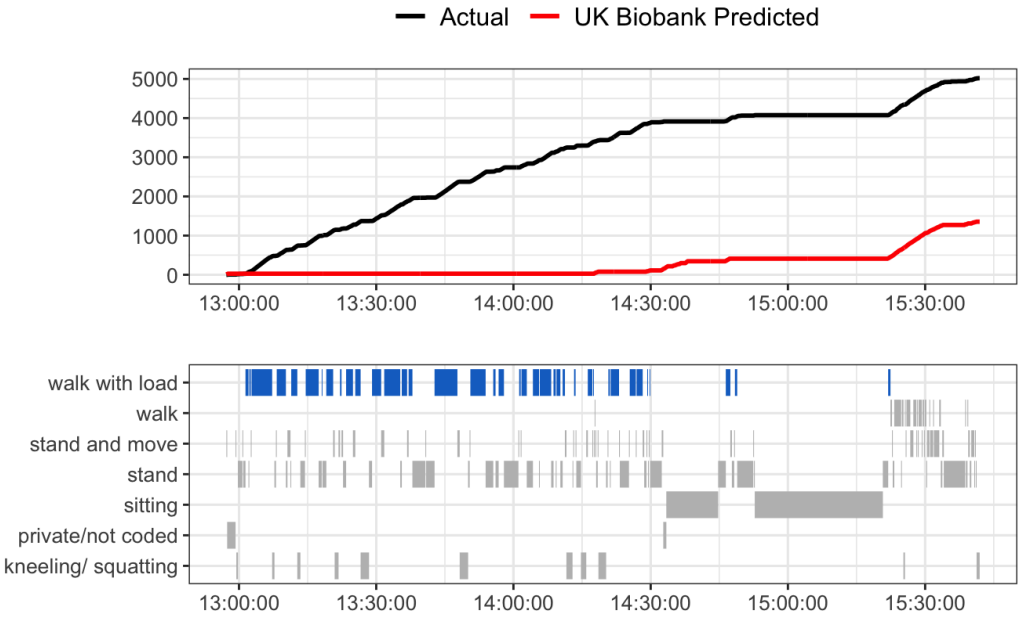


Figure 3. Demonstrates weakness in step prediction during modified walking

We attribute this weakness to the lack of frequency of modified walking within our data, ultimately leading to a LOOCV in which models are heavily impacted by the removal of just one participant. Further work is essential to improve associated step prediction of these modified walking periods, specifically by using more balanced training data that contains more instances of modified walking.

Ultimately, our team has determined preference for the condensed postures set. We believe that among our three algorithms, the condensed postures algorithm satisfies the balance between accurate step prediction and algorithm simplicity. While our Walk/Not walk algorithm is simple, it tends to perform poorly in step prediction. On the other hand, our all postures algorithm is more complex, but tends to perform slightly better than condensed postures in certain scenarios. For example, the all postures set has the lowest LOOCV absolute error per minute of 24.71 steps, compared to 25.85 steps for condensed postures. Moreover, when testing on the OxWalk data, the all postures set has an absolute error per minute of 0.23 steps, whereas the condensed postures set has an absolute error per minute of 0.4 steps. Despite these differences, they are not considerable enough for us to conclude that higher algorithm complexity significantly improves step counting outcomes compared to using condensed postures.

VII. Conclusion

In this study, we aimed to enhance the accuracy of step counting algorithms by refining the UK Biobank algorithm, particularly focusing on periods of modified walking and improving overall step prediction accuracy. By introducing three levels of granularity for posture classification– broad, granular, and condensed– we were able to search for the optimal approach that balanced accuracy and complexity.

Our results showed that classifying more categories generally improved step counting accuracy. Specifically, the condensed posture model hit a strike between simplicity and precision, outperforming the original UK Biobank algorithm in overall step count accuracy when tested on both the ACT24 and OxWalk datasets. Despite the more granular models demonstrating lower errors during cross-validation and testing, their performance during modified walking still showed room for improvement. Our findings suggest that increasing the amount of modified walking in the training data could potentially reduce the error rate, however, obtaining more manually labeled ground truth for such activities is labor-intensive and not always feasible.

In conclusion, our study successfully identified methods to improve the performance of step counting algorithms. The condensed posture model in particular shows a promising direction for future development. Despite the challenges, developing more accurate and reliable step counting algorithms is essential for advancing wearable technology and providing individuals with better insights into their physical activity levels.

VIII. Future Work

In considering the future of our research, there are several areas that are promising for further exploration of step counting algorithms. Firstly, the ACT24 dataset has an extreme imbalance in posture representation, especially certain activities such as running that are very underrepresented. This imbalance poses many challenges for model training because the scarcity of certain postures may hinder the algorithm's ability to accurately classify and predict steps. We believe having more data at our disposal, specifically in postures lacking frequent observation, may improve or change the behavior of our models.

Secondly, while our study focused on a specific set of condensed postures, there remains room for investigating alternative combinations. Exploring a broad range of condensed postures may reveal more effective groupings that balance algorithm accuracy with classification simplicity. Additionally, adopting a more efficient way to create groupings through metrics, rather than relying solely on visual inspection, could result in more precise classifications and improved outcomes.

Finally, we believe other model specifications can be explored. Our approach used a two level algorithm approach: classification and peak detection. Further work on this project may use a one pass approach, where only one step is required to predict steps, or it may use more than two levels. One suggestion we have is creating a 2-step classification then peak detection approach, which is outlined in *Figure 4*.

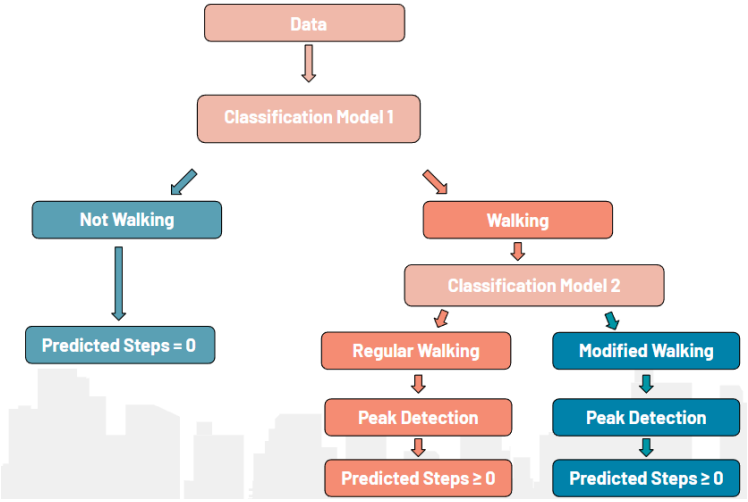


Figure 4. Three Step Algorithm Approach

IX. Appendix

Table 7. LOOCV Walk/Not Walk Classification 10-Second Epoch Confusion Matrix

		Predicted	
		Walk	**Not Walk**
Observed	**Walk**	**24333**	1395
	Not Walk	3082	**3830**

Table 8. LOOCV Condensed Postures Classification 10-Second Epoch Confusion Matrix

		Predicted								
		Asc. Stairs	**Bike**	**Mod. Walk**	**Musc. Stren.**	**No Mvmt.**	**Other Sport Mvmt.**	**Run**	**Stand and Move**	**Walk**
Obs.	**Asc. Stairs**	**3**	0	7	0	117	0	0	15	29
	Bike	0	**0**	45	216	0	0	0	2	0
	Mod. Walk	0	0	**85**	0	336	0	0	34	387
	Musc. Stren.	0	0	0	**0**	59	36	0	0	0
	No Mvmt.	7	0	152	13	**23117**	31	0	1619	233
	Other Sport Mvmt.	0	0	0	0	44	**0**	0	0	0
	Run	0	0	0	0	1	0	**0**	0	0
	Stand and Move	2	0	194	10	1022	0	0	**1408**	275
	Walk	0	0	241	0	250	0	0	283	**2359**

Table 9. LOOCV Granular Classification 10-Second Epoch Confusion Matrix

		Predicted						
		Asc. Stairs	Bike	Desc. Stairs	Kneel/ Squat	Lying	Musc. Stren.	Other Sport Mvmt.
Obs.	Asc. Stairs	**0**	0	1	0	0	0	0
	Bike	0	**0**	0	0	0	215	0
	Desc. Stairs	0	0	**1**	0	0	0	0
	Kneel/ Squat	0	0	0	**43**	0	0	0
	Lying	0	0	0	0	**533**	0	0
	Musc. Stren.	0	0	0	0	0	**0**	14
	Other Sport Mvmt.	0	0	0	0	0	0	**0**
	Run	0	0	0	0	0	0	0
	Sit	0	0	0	36	282	5	1
	Stand	0	0	1	464	0	29	0
	Stand and Move	0	0	2	106	2	16	0
	Stretch	0	0	0	0	0	0	0
	Walk	34	0	279	0	0	1	0
	Walk with Load	0	0	0	274	0	0	0

Cont. Table 9. LOOCV Granular Classification 10-Second Epoch Confusion Matrix

		Predicted						
		Run	**Sit**	**Stand**	**Stand and Move**	**Stretch**	**Walk**	**Walk with Load**
Obs.	**Asc. Stairs**	0	4	4	14	108	36	5
	Bike	0	0	3	0	0	0	44
	Desc. Stairs	0	8	0	4	0	44	2
	Kneel/ Squat	0	57	192	10	21	1	2
	Lying	0	131	0	0	3	0	0
	Musc. Stren.	0	56	0	0	26	0	0
	Other Sport Mvmt.	0	42	0	0	0	0	0
	Run	**0**	0	1	0	0	0	0
	Sit	0	**14795**	883	104	99	63	58
	Stand	0	2708	**3741**	309	158	126	27
	Stand and Move	0	238	1578	**505**	56	258	138
	Stretch	0	279	0	0	**15**	0	0
	Walk	0	109	258	139	14	**1998**	319
	Walk with Load	0	32	46	20	0	297	**113**