Proceedings of the 29th Annual International
Conference of the IEEE EMBS
Cité Internationale, Lyon, France
August 23-26, 2007.

Providing Statistical Measures of Reliability for Body Core Temperature Predictions

Andrei V. Gribok, Mark J. Buller, Reed W. Hoyt, Jaques Reifman

Abstract—**This paper describes the use of a data-driven autoregressive integrated moving average model to predict body core temperature in humans during physical activity. We also propose a bootstrap technique to provide a measure of reliability of such predictions in the form of prediction intervals. We investigate the model's predictive capabilities and associated reliability using two distinct datasets, both obtained in the field under different environmental conditions. One dataset is used to develop the model, and the other one, containing an example of heat illness, is used to test the model. We demonstrate that accurate and reliable predictions of an extreme core temperature value of 39.5 ^{o}C, can be made 20 minutes ahead of time, even when the predictive model is developed on a different individual having core temperatures within healthy physiological limits. This result suggests that data-driven models can be made portable across different core temperature levels and across different individuals. Also, we show that the bootstrap prediction intervals cover the actual core temperature, and that they exhibit intuitively expected behavior as a function of the prediction horizon and the core temperature variability.**

I. Introduction

Heat injury remains a problem for the U.S. armed forces, especially during deployments to localities with very hot climates. From 2003 through 2005, there were 4418 heat injuries and heat illnesses across the services, of which 784 were heat stroke, 3617 were heat exhaustion, and 17 were heat fatalities [1]. To address combat and non-combat injuries of this type, the U.S. Army is developing the Warfighter Physiological Status Monitoring (WPSM) system [2]. The WPSM system consists of wearable sensors that along with decision-support algorithms monitor and predict the physiologic status of soldiers, including body core temperature, a key physiologic indicator of impending heat injuries.

In previous work [3], we developed individual-specific, autoregressive (AR) models for predicting body core temperature during physical activity. We showed that such models, developed based on one individual's data, can be readily ported to other individuals, allowing for the application of individual-specific models without the need to develop and tune core temperature prediction models for each individual. However, in such time- and safety-critical application, it is generally not useful to have the predicted temperature values unless a measure of reliability of the predictions is also provided.

In this paper, we extend our previous work on core temperature predictions to include the application to more realistic field data and, more importantly, to provide a quantitative measure of the reliability of the portable-model predictions by computing prediction intervals (PIs) around the forecasted values. We use data from two field studies involving military activities. The first is used to develop ("train") models, and the second is used to demonstrate that models and PIs can be made portable across studies and across individuals, while providing highly accurate predictions even at limiting thresholds of physiologic health, as one of the subjects in the second study suffered heat exhaustion.

II. Methods and Techniques

A. Data-driven Models

AR models represent a special type of linear data-driven models geared to the prediction of time-series data [4]. In AR modeling, an output signal y_t, at time t, $t=m+1,\ldots,N$, is described as a linear combination of previously observed signals

$$y_t = \sum_{i=1}^{m} b_i y_{t-i} + \varepsilon_t , \quad (1)$$

where b denotes the vector of AR coefficients to be determined, ε_t represents white noise with unknown variance, N denotes the number of training data samples, and m represents the order of the model. If, along with the delayed values of the time series, the delayed values of the noise ε_t are used in (1), the model becomes an autoregressive moving average (ARMA) model. Also, if the time series is

Manuscript received April 02, 2007. This work was supported, in part, by the Combat Casualty Care and the Military Operational Medicine research programs of the U.S. Army Medical Research and Materiel Command, Fort Detrick, Maryland. In collecting the data presented in this manuscript, the investigators adhered to the policies for protection of human subjects as prescribed in Army Regulation 70-25, and the research was conducted in adherence with the provisions of 45 CFR Part 46. The subjects gave their informed consent for the laboratory study after being informed of the purpose, risks, and benefits of the study.

A. Gribok is with the Bioinformatics Cell, Telemedicine and Advanced Technology Research Center (TATRC), U.S. Army Medical Research and Materiel Command (USAMRMC), Frederick, MD 21702, and also with the Nuclear Engineering Department of the University of Tennessee, Knoxville (e-mail: agribok@bioanalysis.org).

M. Buller is a Research Scientist with the U.S. Army Research Institute of Environmental Medicine (USARIEM), Natick, MA 01760 (e-mail: mark.j.buller@us.army.mil).

R. Hoyt is the Chief of Biophysics and Biomedical Modeling Division of USARIEM, Natick, MA 01760 (e-mail: reed.hoyt@us.army.mil).

J. Reifman is a Senior Research Scientist and Director of the Biotechnology High Performance Computing Software Applications Institute, TATRC, USAMRMC, Fort Detrick, MD 21702 (e-mail: Jaques.reifman@us.army.mil).

nonstationary, differentiation is usually applied to restore the stationary nature of the signal, yielding autoregressive integrated moving average (ARIMA) models. These models can be represented as: ARIMA (p, d, q), where *p* and *q* denote the order of the corresponding AR and MA parts, respectively, and *d* denotes the order of differentiation. We use an ARIMA (25, 1, 0) model in this paper.

B. Prediction Intervals

In many safety-critical applications, providing a single-point prediction may not be sufficient. A measure of the reliability of the predictions may be required to assess the uncertainty of the predicted values. To address this problem, an error bound in the form of prediction intervals,

$$\hat{y}_{t+h} \pm z_{\alpha/2}\sqrt{\mathrm{var}(e(t+h))}, \tag{2}$$

can be generated, where $\hat{y}_{t+h}$ is the *h*-step-ahead single-point prediction, $var(e(t+h))$ is the variance of $\hat{y}_{t+h}$, and the prediction factor $z_{\alpha/2}$ is the $\alpha/2$ percentile of a Gaussian distribution [4].

The main problem in estimating PIs is the evaluation of $var(e(t+h))$, for which several approaches have been proposed [4]. In spite of the differences between the different approaches, the fundamental idea behind most of them is that $var(e(t+h))$ can be represented as the sum of two variances

$$\mathrm{var}(e(t+h)) = \mathrm{var}(b(t+h)) + \mathrm{var}(\varepsilon(t+h)), \tag{3}$$

where $var(b(t+h))$ is the variance of the model's parameters from which the confidence intervals are inferred, and $var(\varepsilon(t+h))$ is the variance of the noise in the *h*-step-ahead prediction. These variances are conditioned on the number of training samples *N* and on the prediction horizon *h*.

The bootstrap technique [5] has been widely used to compute variances in data-driven prediction models dealing with independent and identically-distributed (i.i.d.) samples. However, application of the bootstrap technique to time-series data is not as straightforward as it is for i.i.d. problems. This is primarily because the time-dependent structure of the time-series data has to be preserved in any re-sampling procedure, making it difficult to obtain independent replicate samples of the time series [6].

In [7], we developed a bootstrap technique that allows re-sampling for a wide range of autoregressive time-series models, such as ARMA, ARIMA and also for their nonlinear counterparts. The technique explicitly computes both variances, $var(b(t+h))$ and $var(\varepsilon(t+h))$, in (3). The computation of $var(b(t+h))$ relies on the idea of model re-sampling instead of data re-sampling. A population of models is built based on blocks of data that are randomly drawn from the original time series to form an empirical distribution of models. After assembling the empirical distribution of models, re-sampling is performed from this distribution to evaluate $var(b(t+h))$. The steps to compute the distribution of models are as follows (Fig. 1):

1. Randomly draw *B* different bootstrap time-series segments from the original time series. The length and starting time of the segments are selected from a uniform distribution (Step 1).
2. Develop *B* ARIMA models from the *B* bootstrap segments (Step 2).
3. Perform *M* core temperature predictions at time $t+h$ by resampling from the *B* models *M* times, with $M>>B$, generating a distribution of model predictions at time $t+h$ (Step 3).
4. Obtain $var(b(t+h))$ from the distribution of model predictions by applying standard statistical formulation.
5. Obtain $var(b(t+h))$ over the entire time series, for a fixed horizon *h*, by repeating procedures 3 and 4.

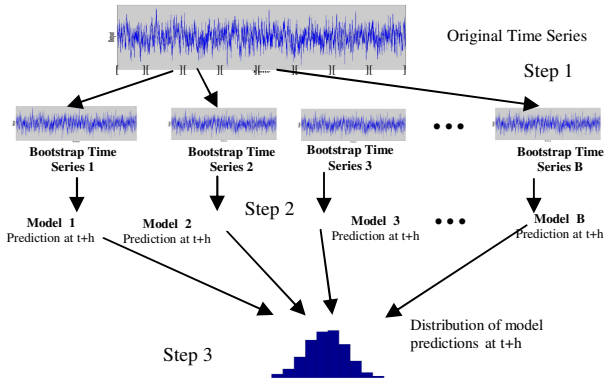

Fig. 1 Bootstrap resampling from models.

Having obtained an estimate of $var(b(t+h))$, we can now use it to compute $var(\varepsilon(t+h))$. Since, in general, $var(\varepsilon(t+h))$ is a nonlinear function of the time-series data, it can be estimated as the output of a feedforward neural network. However, in training the neural network, one cannot simply minimize the sum of the squared errors between the measured outputs and the predictions, as is commonly the case for feedforward neural networks, given that the errors in estimating $var(\varepsilon(t+h))$ cannot be computed directly. Hence, we indirectly train the network to predict $var(\varepsilon(t+h))$ by minimizing the error measure,

$$L = -\sum_{i=1}^{N} \log\left[\frac{1}{\sqrt{2\pi\sigma^2(y_i)}}\exp\left(-\frac{r^2(y_i)}{2\sigma^2(y_i)}\right)\right], \tag{4}$$

expressed as the negative loglikelihood [8], where $\sigma^2(y_i)$ is the variance of the noise in the data at time step *i* and

$$r^2(y_i) = \left\{\left[y_i - E(\hat{y}_i)\right]^2 - \mathrm{var}(b(t+h))\right\}, \tag{5}$$

where *E* is the expectation operator. Notice that we used $var(b(t+h))$ to compute residuals $r^2(y_i)$ in (5), which represent the variance of the noise in the predictions. In other words, we train the neural network with (4) as the cost function, and with

target values estimated from (5). The detailed description of the method is available in [7]. After both terms in (3) are evaluated, we can use (2) to generate PIs. This requires an appropriate selection of the prediction factor $z_{\alpha/2}$. Assuming the distribution of model predictions has a Gaussian distribution, α=5% yields $z_{\alpha/2}$=1.96. However, the assumption of a Gaussian distribution is too restrictive and we opt to use the Camp-Mendel factor [7], which relaxes the normality requirement in favor of a unimodal assumption and gives a confidence factor of 2.98. This confidence factor is used throughout the paper.

III. Results and Discussion

To demonstrate the applicability of the single-point predictions and PIs to field data, we use two datasets: one collected at the Marine Corps base at Quantico, VA, and the other collected at the Joint Readiness Training Center (JRTC), Fort Polk, LA. The Quantico dataset is used to develop ARIMA models, which are subsequently applied to the Fort Polk dataset to produce core temperature predictions and corresponding PIs. For both datasets, core temperatures were collected every minute using an ingested thermometer pill sensor, and the core temperature signals were preprocessed by median and moving average filters to eliminate data artifacts. The data preprocessing was performed off-line, after all temperature profiles became available.

The Quantico dataset [9], collected in July 2001, consists of physiological data collected from eight marine volunteers during a four-day field exercise. Each 10-hour-day involved a 3-mile morning march to a shooting range, followed by day-long exercises and rotations within firing stations, and a march back via the same route in the evening. Subjects wore utility uniforms and, when marching, carried a pack load of 26 kg. A typical temperature profile for the Quantico dataset is presented in Fig. 2.

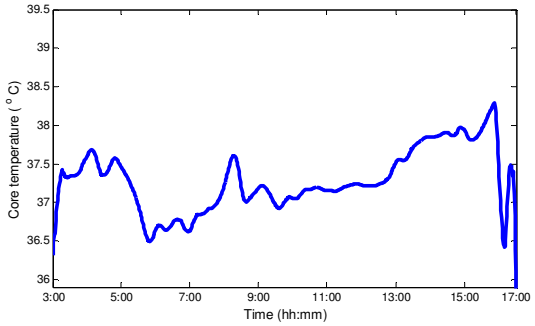

Fig. 2 A typical one-day temperature profile for one subject from the Quantico dataset.

The Fort Polk dataset, collected in August 2001, consists of two subjects, a soldier and a cadet, involved in war games. The soldier and the cadet carried loads of 35 kg and 45 kg, respectively, and both wore utility uniforms with the same thermal resistance as in the Quantico dataset. Figure 3 shows the temperature profiles for the soldier and the cadet. As can be seen in Fig. 3, the cadet's core temperature underwent a sudden increase around 12:30 hrs and reached an extreme value of 39.5 ^{0}C around 13:00 hrs. At this time, the cadet was pulled from the exercise after a member of the JRTC medical staff noticed visible signs of heat exhaustion. This timely medical intervention allowed the cadet to fully recover from the heat illness.

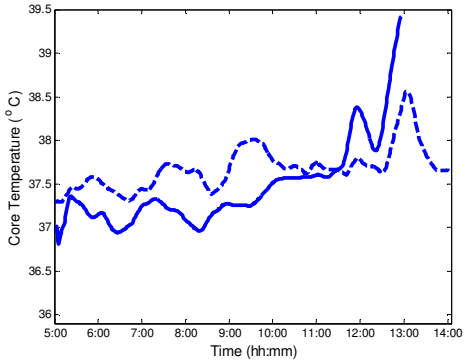

Fig. 3 Temperature profiles for the Fort Polk dataset. Soldier (dashed), cadet (solid).

This dataset is particularly valuable as it presents an opportunity to test the data-driven point and interval estimates at difficult-to-obtain, extreme-temperature conditions.

A randomly selected subject and randomly selected day from the Quantico dataset are used to develop the ARIMA models. After the models are developed, they are applied to the Fort Polk data to produce 20-minute-ahead predictions and corresponding PIs (Fig. 4).

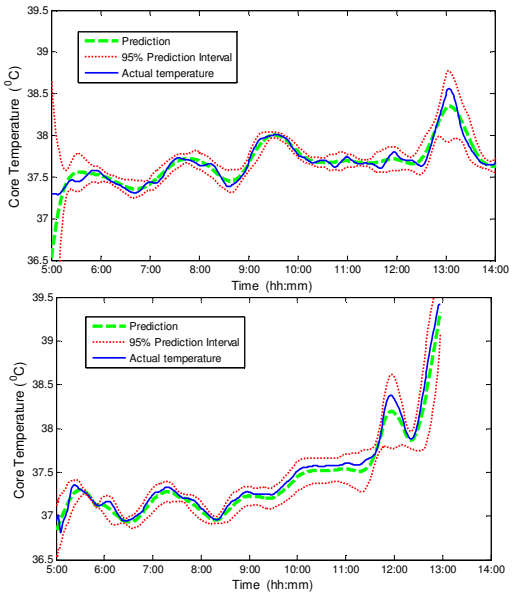

Fig. 4 20-minute-ahead point prediction for the soldier (top panel) and the cadet (bottom panel) along with 95% PIs.

As can be seen, the point predictions are quite accurate and the 95% PIs include the actual measurements for the whole

duration of the test data. An important point is that the ARIMA model is able to predict at 12:40 hrs that the cadet's core temperature will reach the dangerous limit of 39.5 ^{0}C at 13:00 hrs, thus providing sufficient time for preventive medical intervention. These results are especially encouraging because the Quantico dataset, used for training, does not contain any temperature values near 39.5 ^{0}C, suggesting that predictive data-driven models can be developed on individuals having core temperature within healthy physiological limits to predict other individuals at extreme temperature conditions. Hence, making those models portable across different individuals and across different core temperature levels.

Notice that the PIs are also larger in the regions with rapid changes of the core temperature signal, which is intuitively correct as there is more uncertainty in predictions in such regions.

We also test the behavior of the PIs as a function of the prediction horizon. Figures 5 and 6 show the PIs for 2- and 30-minute-ahead-prediction horizons, respectively, for the cadet. As can be seen, the PIs become wider as the prediction horizon increases, which is intuitively correct as we expect the confidence in the predictions, represented by the width of the PIs, to deteriorate as the prediction horizon increases.

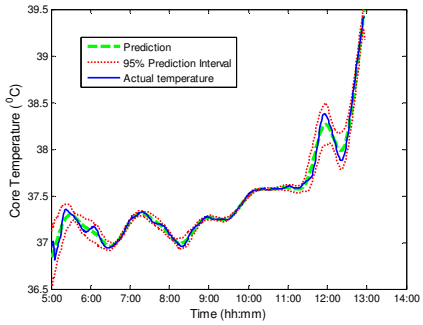

Fig. 5 2-minute-ahead point predictions along with corresponding 95% PIs for the cadet.

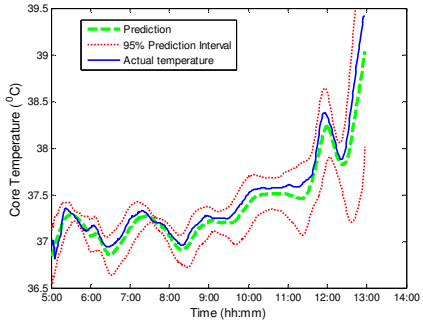

Fig. 6 30-minute-ahead point predictions along with corresponding 95% PIs for the cadet.

The success of ARIMA modeling for core temperature predictions is based on the property of the human body to absorb a significant amount of energy without changing its core temperature. This is due to the fact that water, the main component of the human body, has one of the highest specific heat capacities among all substances. It means that the human body is a very "heavy" object in thermal sense, possessing a large amount of thermal inertia. The significant thermal inertia causes a high degree of correlation among consecutive core temperature measurements, allowing the ARIMA model, which relies on such correlations, to produce highly accurate and statistically reliable predictions.

IV. Conclusions

Accurate single-point predictions of thermal status, i.e., core temperature, can be produced by data-driven, ARIMA models even when such models are developed on datasets that do not contain extreme core temperature variations.

The bootstrap prediction intervals can be placed around single-point predictions, thereby providing a measure of confidence for the core temperature predictions. The proposed prediction intervals exhibit an intuitively expected behavior as they widen with increased prediction horizons and in regions where the core temperature undergoes rapid changes.

The proposed algorithm for core temperature prediction is currently being implemented as a part of the WPSM system. It will undergo extensive field studies and, if proven successful, could also be used to predict impending heat injuries during civilian activities, such as sport and strenuous work.

Disclaimer

The opinions or assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the U.S. Army or of the U.S. Department of Defense.

References

[1] *Memorandum for Heat Injury Prevention Program,* Department of the Army, Office of the Surgeon General, 5109 Leesburg Pike, Falls Church, VA 22041-3258, 8 April 2005, 17 April 2006, http://chppm-www.apgea.army.mil/heat/ , accessed on 11/01/2006.

[2] R.W. Hoyt, J. Reifman, T. S. Coster, and M. J. Buller, "Combat medical informatics: present and future," in *Proceedings of the AMIA Annual Symposium*, 2002, pp. 335-339.

[3] A. Gribok, T. McKenna, and J. Reifman, "Regularization of Body Core Temperature Prediction during Physical Activity," in *Proceedings of the IEEE International Conference of the Engineering in Medicine and Biology Society. Engineering Revolution In BioMedicine*, 2006, pp. 459-463.

[4] C. Chatfield, *Time-Series Forecasting*. Boca Raton, London, New York, Washington DC: Chapman and Hall/CRC Press, 2000.

[5] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. London: Chapman and Hall, 1993.

[6] D. N. Politis, "The impact of bootstrap methods on time-series analysis," *Statistical Science*, vol. 18, pp. 219-230, 2003.

[7] N. Oleng, A. Gribok, and J. Reifman, "Error bounds for data-driven models of dynamical systems," *Computers in Biology and Medicine*, vol. 37, pp. 670-679, 2007.

[8] T. Heskes, "Practical confidence and prediction intervals," in *Proceedings of NIPS,* vol. 9, M. Mozer, M. Jordan, and T. Petsche, Ed., MIT Press, 1997, pp. 176-182.

[9] M. Yokota, L. Berglund, W. R. Santee, M. J. Buller, and R. W. Hoyt, "Modeling Physiologic Responses to Military Scenarios: Initial Core Temperature and Downhill Work," *Aviation, Space, and Environmental Medicine,* vol. 76, pp. 475-480, 2005.