PharmaSUG 2018 – Paper DV-11

Processing and Cleaning Streaming Data in SAS

Mehmet Kocak, The University of Tennessee Health Science Center; Rebecca Krukowski, Gerald Wayne Talcott, The University of Virginia

ABSTRACT

The streaming data has become part of everyone's daily life with the common use of smart devices. It is also part of almost every clinical research protocol with the increasing use of smart monitoring systems such as e-scales, wearable blood pressure or heart rate monitors, etc. Although capturing and storing such data is a growing challenge, making such data available for statistical analysis is another challenge for statisticians and computer programmers. In this research, we propose an algorithm aiming to capturing the true profile in a given streaming data with a specific application to BodyTrace weight data captured in the Fit Blue study. We show that our approach is highly sensitive in identifying the correct profile and highly specific in cleaning the correct profile from other mixing profiles. We have implemented the proposed approach in a SAS macro program named %TPF (**T**rue**P**rofile**F**inder).

INTRODUCTION

With the widespread availability of wireless internet connection as well as other signals such as GPS signals, almost everyone is part of generating a form of streaming data, be it locational data, be it self-entered and automatically captured health-related or behavioral data. Such data is continuously captured and mined especially for marketing purposes to generate potential customer profiles by tapping into the multi-dimensional data streams.

Such data streams are also part of almost every clinical research protocol with the increasing use of smart monitoring systems such as e-scales, wearable blood pressure or heart rate monitors, etc. (Steinberg et al., 2015; Mao et al., 2017), as these devices present an easy access for the investigators to reach out to their study participants, or vice versa, without much difficulty for more granule and real-time data collection as well as for faster interventions in case of disease worsening.

Such multi-dimensional data flowing continuously and growing exponentially over time requires a collaborative efforts among investigators, computer scientists, and statisticians as to what to keep from the stream, when to keep them, and how to store and access them. The issue of storing such data aside, some processes require real-time intervention based on the acquired data to that time point, especially in patient monitoring. Detecting a change-point, where a change beyond a tolerable threshold may indicate worsening of patient status, requiring immediate medical intervention, may give a much needed window of opportunity for the treating physician to intervene and develop an immediate action plan for disease management on a real-time basis.

Although capturing and storing such data is a growing challenge, making such data available for statistical analysis is another challenge for statisticians and computer programmers (Yang & Wu, 2006; Zhou et al., 2009; Zikopoulos & Eaton, 2011). In this, while signal to noise ratio is an existing problem, it is also possible that the streaming data may be a mixture of signals beyond the desired signal alone, and it requires a process to detect the true signal on a real-time basis as well as in post-hoc data processing (O'callaghan et al., 2002; Babcock et al., 2002; Jin & Agrawal, 2003; Kanagal & Deshpande, 2008; Ross & Wing, 2016), so that the actions, interventions, and inference can be made based on the true signal.

In this research, we propose an algorithm aiming to capturing the true profile in a given streaming data from individuals; our proposal can also incorporate the true signal values called 'clinic values', which is considered 'certainly correct signal' as they are acquired in a clinic visit or at lab-environment. We apply our proposed method to the BodyTrace weight data captured in the Dissemination of the Look AHEAD Weight Management Treatment in the Military (Fit Blue) study (Krukowski et al., 2015; Maclin-Akinyemi et al., 2017). We show that our approach is highly sensitive in identifying the correct profile and highly specific in cleaning the correct profile from other mixing profiles. We have implemented the proposed approach in a SAS macro program named %TPF (**T**rue**P**rofile**F**inder).

We present the algorithm of our proposal and SAS %TPF macro program in Section 2, followed by the real-data application in Section 3. We provide our conclusions and future research plans in Section 4.

SECTION-2: THE ALGORITHM

In a given streaming data, to identify the true profile, our %TPF macro program utilizes the following steps:

1. Sort the signal chronologically;
2. Determine a window size such as 10, which will split the observations in the bins of size 10;
3. Determine a reasonable 'similarity' band such as 5, which will indicate two signals that are within 5 units of each other be considered 'coming from the same process';
4. Starting from the first bin, assign the weight of 1.0 to each signal values in the selected bin; if there are 'clinic' signals which we are certain to come from the true signal, assign a higher weight such as 5 or 10.
5. Fit a desired level of a polynomial regression; this can be linear, quadratic, cubic, etc., and compute the absolute residuals for the selected bin;
6. If the residual of a signal is within the chosen 'similarity band', call it 'true signal' and assign the weight of 1.0; otherwise, call it 'not true signal' and assign the weight of 0.0;
7. Move to the next bin, keeping all the 'true signals' identified in the decision process also in your current active data for decisions. This will allow a possibility that a given signal identified as 'true' may be called as 'not true' signal based on the future data;
8. Apply the steps 4-7 iteratively to cover all the bins.

To monitor the process, the user can generate graphical tools as well.

We present the SAS %TPF macro program below:

```
*-------------------------------------------------------------------------------
* Mehmet Kocak 01/01/2018
* This macro program performs a data cleaning operations to identify the true signal
  in a streaming data. The procedure can utilize 'clinic' measurements which are
  considered 'absolutely correct'.
The user need to provide values to the following parameters:
 DATA: Input dataset
 TVAR: Time variable
 YVAR: Signal variable
 WSIZE: Window Size. This is the size of the moving window detecting new 'correct
        profile' measurements and it starts from the first observation.
 MGROUP: Measurement Group. This variable identifies a given measurement as 'clinic'
        or 'data stream' measurement. This is critical to identify as these two
        measurement types will be weighed differently in the correct-profile
        detection process. If you don't have any clinic data, then you still need to
        create an MGROUP variable and assign the value of 0 to all measurements
        indicating no clinic measurement.
 MTYPE: The code for the 'clinic' measurements in MGROUP variable. Default is 1.
 AWEIGHT:    The weight of the actual 'clinic' measurements. Default is 10.
 MODELCHOICE: The choice of the polynomial model. The variable TT will be created
              during the operation so you simply specify the model using variable
              TT.
 LOWERBND: The lower bound of measurements below which all measurements are
           considered not belonging to the correct profile without modeling.
 UPPERBND: The upper bound of measurements above which all measurements are
           considered not belonging to the correct profile without modeling.
 PREDBAND: The prediction band around the model prediction beyond which all
           measurements will be considered not belonging to the correct profile.
 TRUEPROFILE:The variable indicating true profile if available. If not, a new code
             will be added.
```

```
 OUTNAME: The output dataset name to retain the original data as well as the final
          profile calls in a variable called FINALPROFILE.
*-------------------------------------------------------------------------------------;

%macro TPF(data=, tvar=, yvar=, wsize=7, mgroup=dtype, mtype=1, aweight=10,
modelchoice=tt|tt, lowerbnd=100, upperbnd=300, predband=11, trueprofile=,
outname=mydata); options nonotes;
options nonotes;
proc format;
value profile 0='Actual Data' 1='Correct Profile' 2='Incorrect Profile' 9='Unknown';
run;
proc sort data=&data; by &tvar; run;
data indata0; set &data;
if &mgroup=&mtype; run; data indata1; set &data;
if &mgroup^=&mtype; timid=_n_; run;
data indata; set indata0 indata1;
%if &trueprofile=%str() %then %do; true_profile=9;
if &mgroup=&mtype then true_profile=0; %end;
%else %do; true_profile=&trueprofile; %end;
run;
proc sql; drop table indata0, indata1; quit;
proc sql noprint; select distinct count(*) into :nofrecords
from indata where timid^=.; quit;
proc sql noprint; select distinct floor(count(*)/&wsize) into :binsize
from indata where timid^=.; quit;
data indata; set indata; recid=_n_; yy=&yvar; tt=&tvar;
if  &mgroup=&mtype then do; cweight=&aweight; finalprofile=0; end;
else if  yy<=&lowerbnd or yy>=&upperbnd then do; cweight=0; finalprofile=3; end;
else if timid>1 and timid<=&wsize then do; cweight=1; finalprofile=1; end;
if finalprofile in (0 1) then include=1; run;
proc glm data=indata plots=none; where include=1; weight cweight; model
yy=&modelchoice; output out=preddata p=predicted; run;
proc sql; create table dataupdate as select distinct recid, predicted
from preddata order by recid; quit;
data indata; merge indata dataupdate; by recid;
if include=1 and finalprofile not in (0 3) then do;
resid=round(abs(yy-predicted),0.1);
       if resid>0 and resid<=&predband then do; finalprofile=1; cweight=1; end;
       else if resid>&predband then do; finalprofile=2; cweight=0; end;
end; run;
data indata; set indata; drop predicted resid; run;
proc sql; drop table preddata, dataupdate; quit;
%do i=1 %to &binsize;
data indata; set indata;
if finalprofile in (0 1) or (timid>=%sysevalf(&i*&wsize-&wsize) and
timid<=%sysevalf(&i*&wsize+&wsize)) then include=1;
%if %sysevalf(&binsize+0)=%sysevalf(&i+0) %then  %do; include=1; %end; run;
proc glm data=indata plots=none; where include=1; weight cweight; model
yy=&modelchoice; output out=preddata p=predicted; run;
proc sql; create table dataupdate as select distinct recid, predicted
from preddata order by recid; quit;
data indata; merge indata dataupdate; by recid;
if include=1 and finalprofile not in (0 3) then do;
resid=round(abs(yy-predicted),0.1);
       if resid>0 and resid<=&predband then do; finalprofile=1; cweight=1; end;
       else if resid>&predband then do; finalprofile=2; cweight=0; end;
end; run;
data indata; set indata; drop predicted resid; run;
proc sql; drop table preddata, dataupdate; quit;
%end;
options notes;
proc sql; create table &outname as select distinct recid, true_profile, &mgroup, tt
as &tvar, yy as &yvar, finalprofile from indata; quit;
```

```
proc sql; drop table indata; quit;
proc tabulate data=&outname; class true_profile finalprofile;
label true_profile='Correct Profiles' finalprofile='Predicted Profiles';
table true_profile, finalprofile*n=''; format true_profile finalprofile profile.;
run;
%mend TPF;

/*
***    A sample call of the TPF Macro: ***

*** Sample Data ****;
data sampledata; call streaminit(23134);
time=0; weight=200; dtype=1; wgth=10; correctprofile=0; output;
do time=1 to 19 by 2; weight=200-1*time+rand('normal')*5; dtype=2; wgth=1;
correctprofile=1; output; end;
do time=2 to 18 by 2; weight=180-1*time+rand('normal')*5; dtype=2; wgth=1;
correctprofile=2; output; end;
time=20; weight=180; dtype=1; wgth=10; correctprofile=0; output;
do time=21 to 39 by 2; weight=180-1*(time-20)+rand('normal')*5; dtype=2; wgth=1;
correctprofile=1; output; end;
do time=20 to 38 by 2; weight=160-1*(time-20)+rand('normal')*5; dtype=2; wgth=1;
correctprofile=2; output; end;
time=40; weight=160; dtype=1; wgth=10; correctprofile=0; output;
run;

%tpf(data=sampledata, tvar=time, yvar=weight, wsize=5, mgroup=dtype, mtype=1,
aweight=10, modelchoice=tt, lowerbnd=100, upperbnd=250, predband=10,
trueprofile=correctprofile, outname=sampledata_profile); */
```

SECTION 3: APPLICATION TO REAL-DATA

To implement our proposed approach using the SAS %TPF macro program, we will use the BodyTrace weight data from the Dissemination of the Look AHEAD Weight Management Treatment in the Military (Fit Blue) Study (Krukowski et al., 2015; Maclin-Akinyemi et al., 2017). The Fit Blue study (http://www.uthsc.edu/fitblue/info.php) tailors the evidence-based Look AHEAD weight loss intervention, which has been validated in a civilian population, to the active duty military population, specifically the US Air Force. The study compares outcomes from two groups: one group receives a phone-based intensive counselor-directed weight loss intervention, while the other group receives a phone-based self-directed weight loss intervention. The study enrolled over 248 active duty military personnel stationed at Joint Base San Antonio – Lackland, and the study participants were given e-scales through which they measured their weights any time and their weight data were wirelessly transmitted to the BodyTrace database.

From 248 participants in the Fit Blue study, we captured 55,788 instances of weight measurements (~225 per participant). Naturally, not only the participant, but other family members of the participants, or their guests, even pets, may also step on the e-scale, which then transmits a new weight measurement to the database, generating ‘incorrect’ profiles. Because of this, before such data can be used in statistical analyses for publications, a careful data cleaning process should be undertaken to detect the true weight signals. Figures 1 and 2 below depict good examples of mixed signals from the BodyTrace weight data:

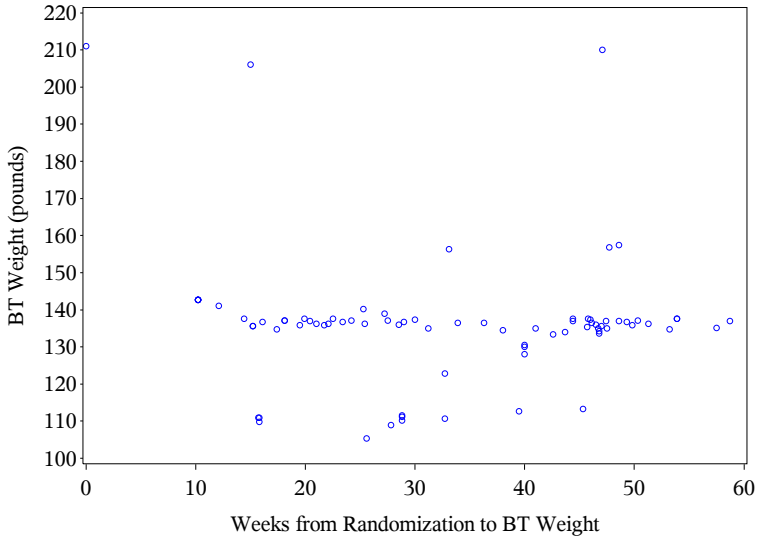

Figure 1. Example-1 of mixture signals from BodyTrace weight data

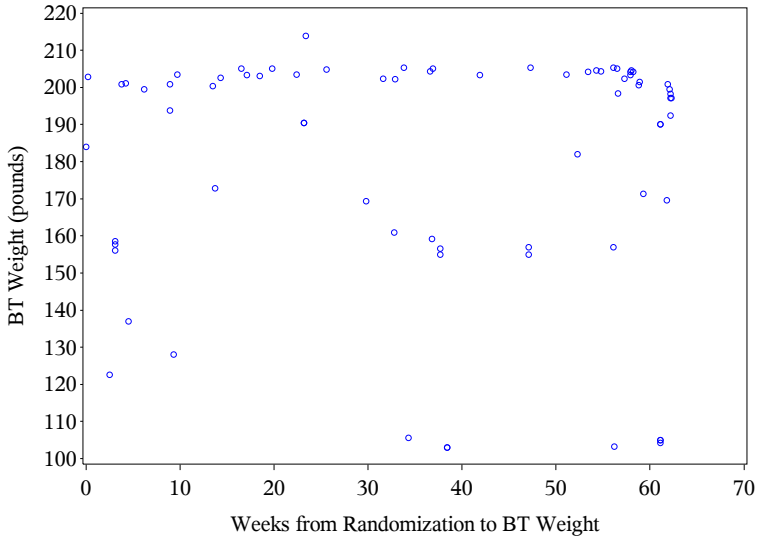

Figure 2. Example-2 of mixture signals from BodyTrace weight data

In this cleaning process, we utilized the clinic weights from Baseline, 4-month, and 12-month visits as much as possible. We processed each profile manually one-by-one, and classified each weight measurement as 'from the true signal' vs. 'not from the true signal', and the final calls were reviewed and approved by the study team; this manual cleaning will serve as our GOLD STANDARD to assess the performance of the SAS %TPF macro program.

Figure 3 illustrates the result of an application of %TPF macro program to a selected participant:

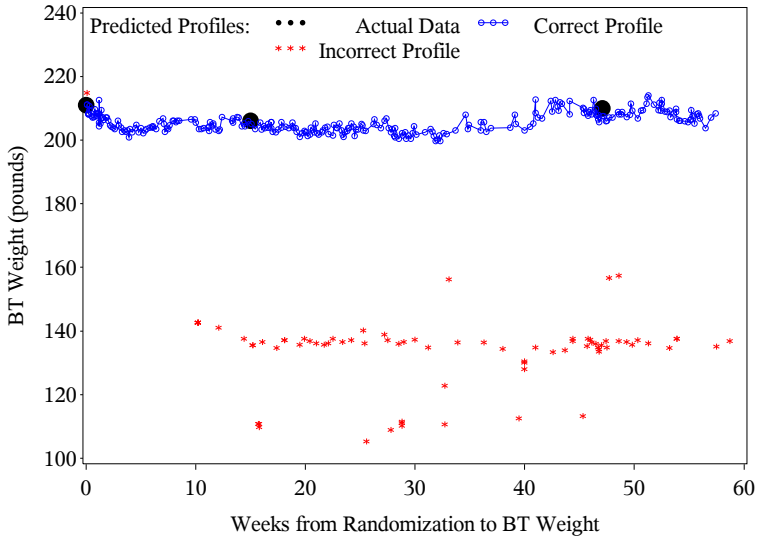

Figure 3. An application of %TPF macro program to a selected participant

We applied the SAS %TPF macro program with the following specifications to the Fit Blue BodyTrace weight data:

- Model Choice: Linear, Quadratic, Cubic, and Quartic
- Data-bin size: 5, 10, 20
- Clinic data weight: 5, 10

Once the profile calls are made based on the SAS %TPF macro processing, we have computed the sensitivity, specificity, and percentage of true calls under each scenario considering our 'manual cleaning' calls as GOLD STANDARD.

Figure 4 and Table 1 below illustrate the results from these 24 scenarios tested.

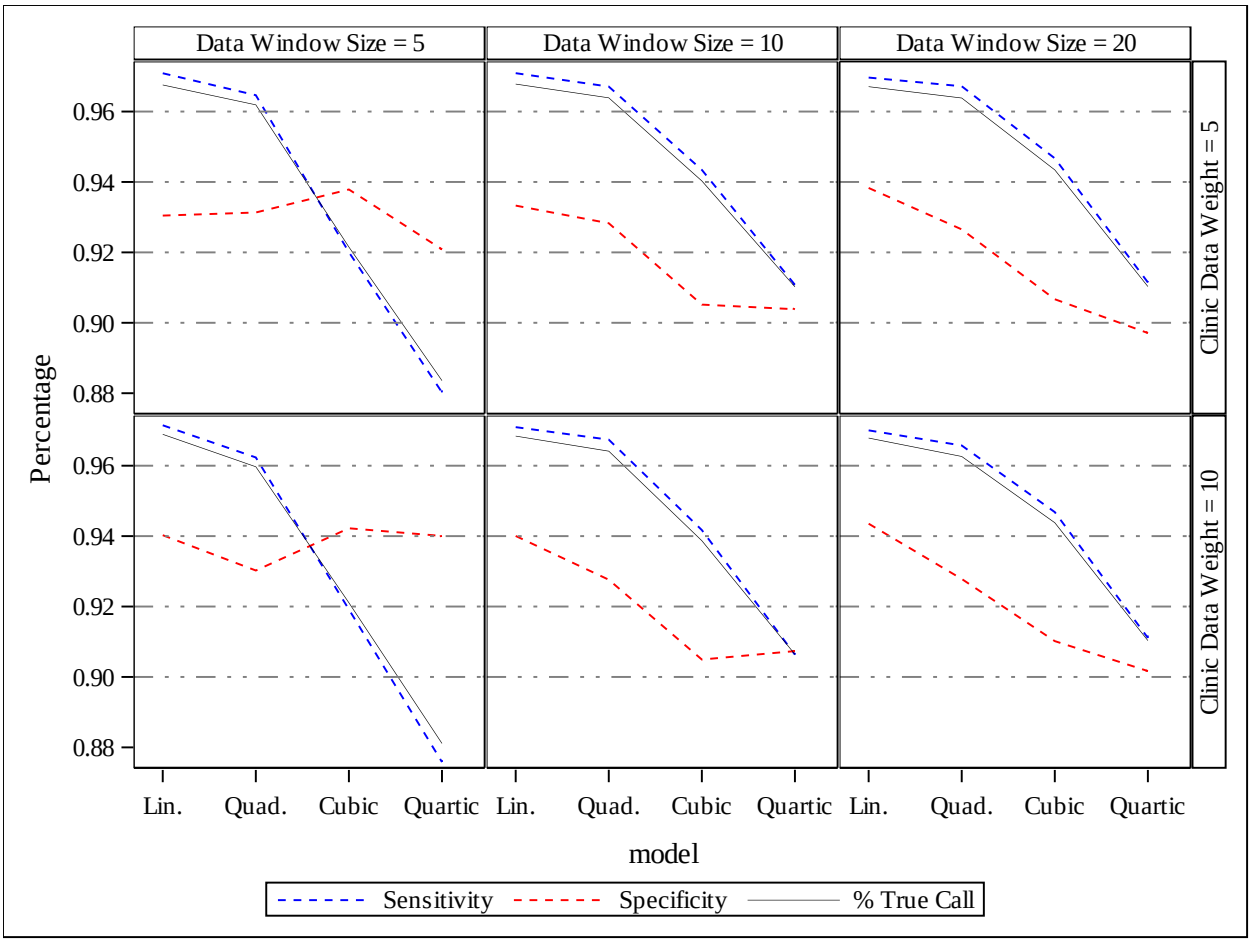

Figure 4. Sensitivity, Specificity, and True Call Percentage from 24 scenarios tested

Table 1. Sensitivity, Specificity, and True Call Percentage from 24 scenarios tested (sorted by Percent True Call in a descending order)

Window Size	Clinic Data Weight	Model	Sensitivity	Specificity	Percent True Call
5	10	Lin.	0.971	0.94	0.969
10	10	Lin.	0.971	0.94	0.968
20	10	Lin.	0.97	0.944	0.968
10	5	Lin.	0.971	0.933	0.968
5	5	Lin.	0.971	0.93	0.968
20	5	Lin.	0.97	0.938	0.967
10	10	Quad.	0.967	0.928	0.964
10	5	Quad.	0.967	0.928	0.964
20	5	Quad.	0.967	0.927	0.964
20	10	Quad.	0.966	0.928	0.963
5	5	Quad.	0.965	0.931	0.962
5	10	Quad.	0.962	0.93	0.96
20	10	Cubic	0.947	0.91	0.944
20	5	Cubic	0.947	0.907	0.943
10	5	Cubic	0.943	0.905	0.94
10	10	Cubic	0.942	0.905	0.939
5	5	Cubic	0.92	0.938	0.922
5	10	Cubic	0.919	0.942	0.921

Window Size	Clinic Data Weight	Model	Sensitivity	Specificity	Percent True Call
20	10	Quartic	0.911	0.902	0.91
20	5	Quartic	0.911	0.897	0.91
10	5	Quartic	0.911	0.904	0.91
10	10	Quartic	0.906	0.907	0.906
5	5	Quartic	0.88	0.921	0.884
5	10	Quartic	0.876	0.94	0.881

We conclude that, in this particular application to Fit Blue BodyTrace data, linear fit seems to outperform the other polynomial model selection. Assigning higher weights to the available clinic weights results in a small (mostly negligible) increase in performance of the profile cleaning approach. Similarly, larger bin-size results in a small increase in performance as well.

CONCLUSION

With this application, we have shown that our proposed correct profile detection algorithm, SAS %TPF macro, is an efficient way of performing an initial data cleaning step for BodyTrace data in particular or any other streaming data in general. For such streaming data, a linear fit seems to work much better; this is not surprising considering that the change within a short window would be detected more easily by a linear fit than higher order polynomial fits. Here, the user must be cognizant of the fact that the bin-size should accommodate a choice of polynomial fit. For example, if the user chooses a 5-degree polynomial, the bin-size must be greater than 5 so that such fit can be possible.

In this particular application, our main objective was to identify the true signal which comes from only one source. This can be expanded to more than one source of streaming data; for example, we can think of a weight-loss study which enrolls couples, leading to at least two true signals if only one e-scale is used.

Also, in our current approach, we use a linear regression modeling, which can be expanded to other non-linear approaches; for example, Logistic or Gompertz growth curves can be used as the base model to detect the signal in each choice of time-window and updated iteratively.

In addition, in our current approach, we only utilize the measurement time as our only predictor of the signal; this can be expanded to other participant characteristics which may or may not change within subject.

REFERENCES

Steinberg, D. M., Bennett, G. G., Askew, S., & Tate, D. F. (2015). Weighing every day matters: daily weighing improves weight loss and adoption of weight control behaviors. *Journal of the Academy of Nutrition and Dietetics*, *115*(4), 511-518.

Mao, A. Y., Chen, C., Magana, C., Barajas, K. C., & Olayiwola, J. N. (2017). A mobile phone-based health coaching intervention for weight loss and blood pressure reduction in a national payer population: A retrospective study. *JMIR mHealth and uHealth*, *5*(6).

Yang, Q., & Wu, X. (2006). 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making*, *5*(04), 597-604.

Zhou, A. Y., Jin, C. Q., Wang, G. R., & Li, J. Z. (2009). A survey on the management of uncertain data. *Chinese Journal of Computers*, *32*(1), 1-16.

Zikopoulos, P., & Eaton, C. (2011). *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media.

O'callaghan, L., Mishra, N., Meyerson, A., Guha, S., & Motwani, R. (2002). Streaming-data algorithms for high-quality clustering. In *Data Engineering, 2002. Proceedings. 18th International Conference on* (pp. 685-694). IEEE.

Babcock, B., Datar, M., & Motwani, R. (2002, January). Sampling from a moving window over streaming data. In *Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete algorithms* (pp. 633-634). Society for Industrial and Applied Mathematics.

Jin, R., & Agrawal, G. (2003, August). Efficient decision tree construction on streaming data. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 571-576). ACM.

Kanagal, B., & Deshpande, A. (2008, April). Online filtering, smoothing and probabilistic modeling of streaming data. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on* (pp. 1160-1169). IEEE.

Ross, K. M., & Wing, R. R. (2016). Concordance of in-home 'smart'scale measurement with body weight measured in-person. *Obesity science & practice*, *2*(2), 224-228.

Krukowski, R. A., Hare, M. E., Talcott, G. W., Johnson, K. C., Richey, P. A., Kocak, M., ... & Klesges, R. C. (2015). Dissemination of the Look AHEAD intensive lifestyle intervention in the United States Air Force: Study rationale, design and methods. *Contemporary clinical trials*, *40*, 232-239.

Maclin-Akinyemi, C., Krukowski, R. A., Kocak, M., Talcott, G. W., Beauvais, A., & Klesges, R. C. (2017). Motivations for Weight Loss Among Active Duty Military Personnel. *Military medicine*, *182*(9-10), e1816-e1823.

ACKNOWLEDGMENTS

- The research represents a Collaborative Research and Development Agreement with the United States Air Force (CRADA #13-168-SG-C13001).
- The study was funded by the National Institute of Diabetes and Digestive and Kidney Diseases (RO1 DK097158, PI: Krukowski, Klesges).
- The opinions expressed in this document are solely those of the authors and do not represent an endorsement by or the views of the United States Air Force, the Department of Defense, or the United States Government.
- The views of BodyTrace TM are not necessarily the official views of, or endorsed by, the U.S. Government, the Department of Defense, or the Department of the Air Force. No Federal endorsement of BodyTrace TM is intended.
- We gratefully acknowledge the partnership with BodyTrace TM.
- We would like to thank the participants and the research team for their dedication to the research. This paragraph uses the PaperBody style.
- The views expressed in this talk are those of the authors and do not reflect the official policy or position of the Department of the Air Force, Department of Defense, or the U.S. Government.

RECOMMENDED READING

- *Base SAS® Procedures Guide*
- *SAS Macro Language Magic: Discovering Advanced Techniques by Robert Virgile*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Mehmet Kocak, Ph.D.
Associate Professor of Biostatistics

Department of Preventive Medicine
The University of Tennessee Health Science Center
Memphis, TN 38103
(901) 448-2937
Mkocak1@uthsc.edu