KU KDM 2019
Konferencja Użytkowników Komputerów Dużej Mocy
Zakopane, 6-8 marca 2019 r.

On Approach to Analysis of Scarce Poor Quality Data Supplemented by Randomly Generated One

O koncepcji analizy mało-licznych danych eksperymentalnych o niskiej jakości wspomaganych przez losową symulację

Janusz Orkisz, Jadwiga Zaborska
Institute for Computational Civil Engineering
Cracow University of Technology

INTRODUCTION

YEAR AGO

SAME PROBLEM

DIFFERENT SOLUTION METHOD

GENERAL DATA CHARACTERISTICS

data amount	data	quality
	good	poor
numerous	very good	~ OK
too little	~ OK	hopeless?

VIEW POINTS

PESSIMISTS : HOPELESS

OPTIMISTS : ? - TRY !

SOLUTION APPROACH

LAST YEAR: POOR DATA + HEURISTICS

PROBLEM: $(\bar{u},\sigma)_4 \div (\bar{u},\sigma)_{100}$

THIS YEAR: POOR DATA+RANDOM PSEUDO DATA (NO MODEL CASE)

PROVIDES:

- $(\bar{u},\sigma)_4$ AND $(\bar{u},\sigma)_{100}$ RELATION
- NEW SOLUTION APPROACH

CONTENTS

PHYSICAL PROBLEM CONSIDERED

MEASUREMENTS

- MUSCLES (EXTENSORS, FLEXORS) STRENGTH
- USE TRAINING DEVICE „ATLAS”
- OBJECTIVE: EVALUATION OF TREATMENT (TRAINING) EFFECT
- DATA CHARACTERISTICS AND RESULTS

ANALYSIS

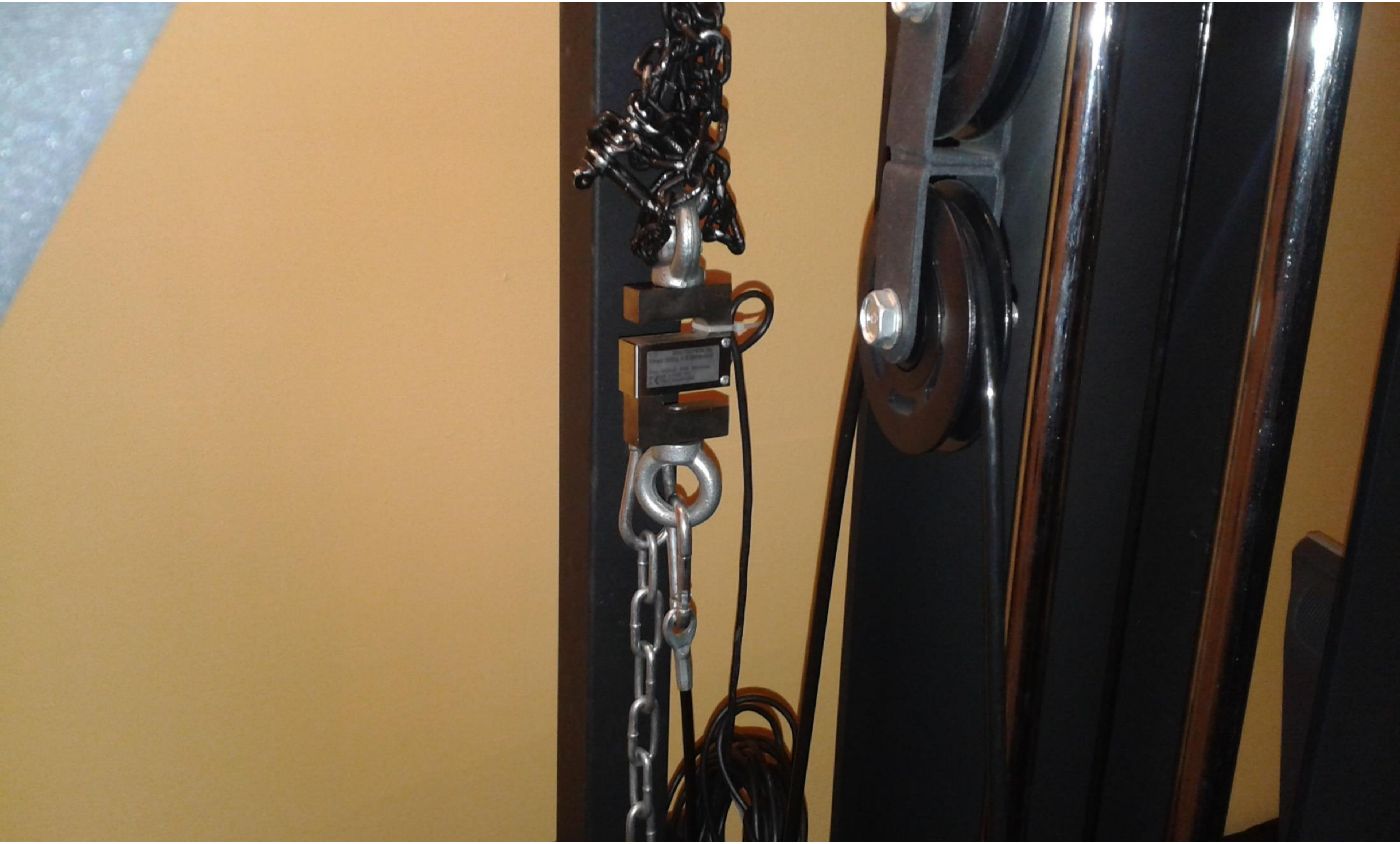
= STANDARD

= INNOVATIVE HEURISTIC ERROR FUNCTIONALS

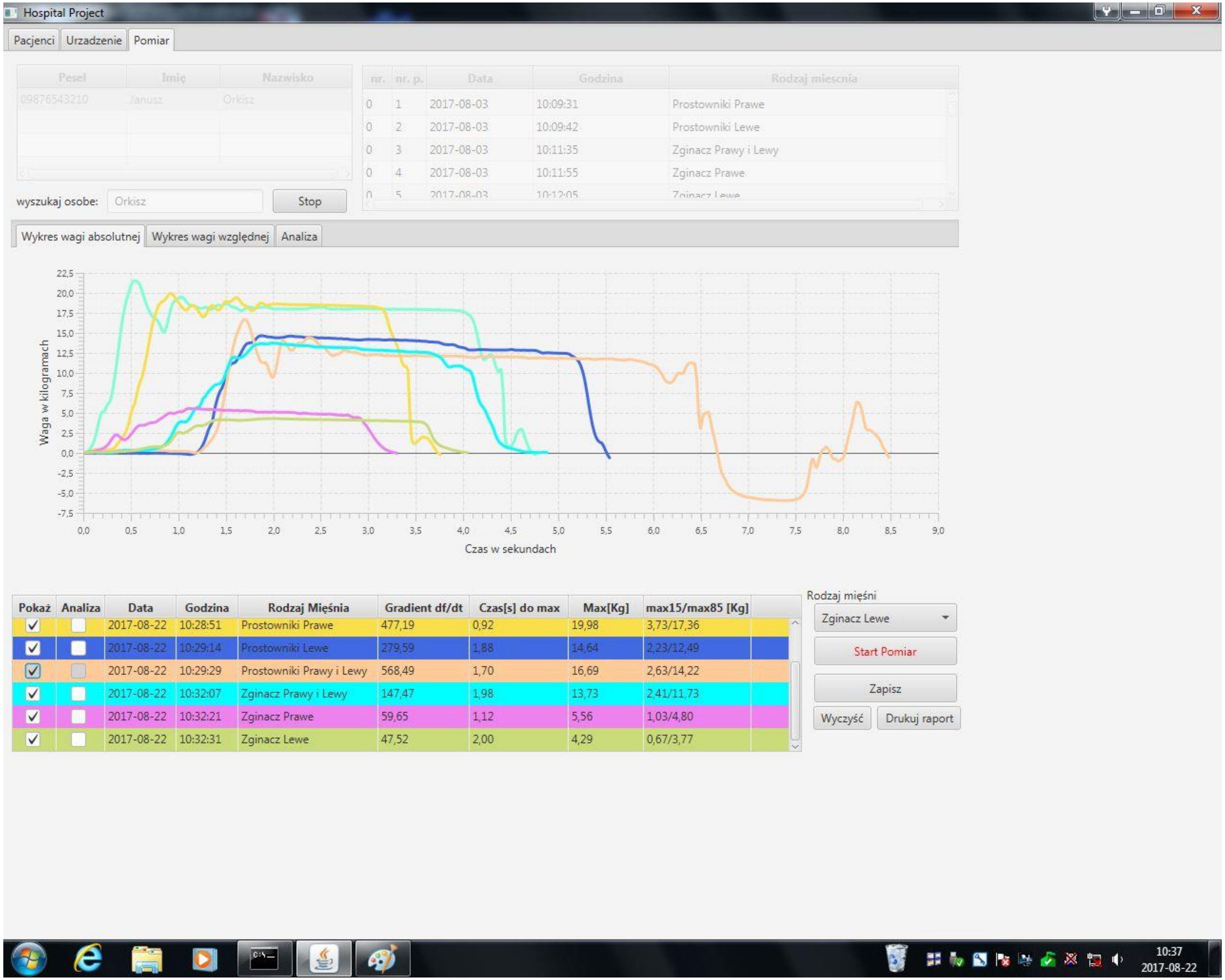
= SUPPORT OF RANDOM PSEUDO DATA

TRUE MEASUREMENTS - ATLAS

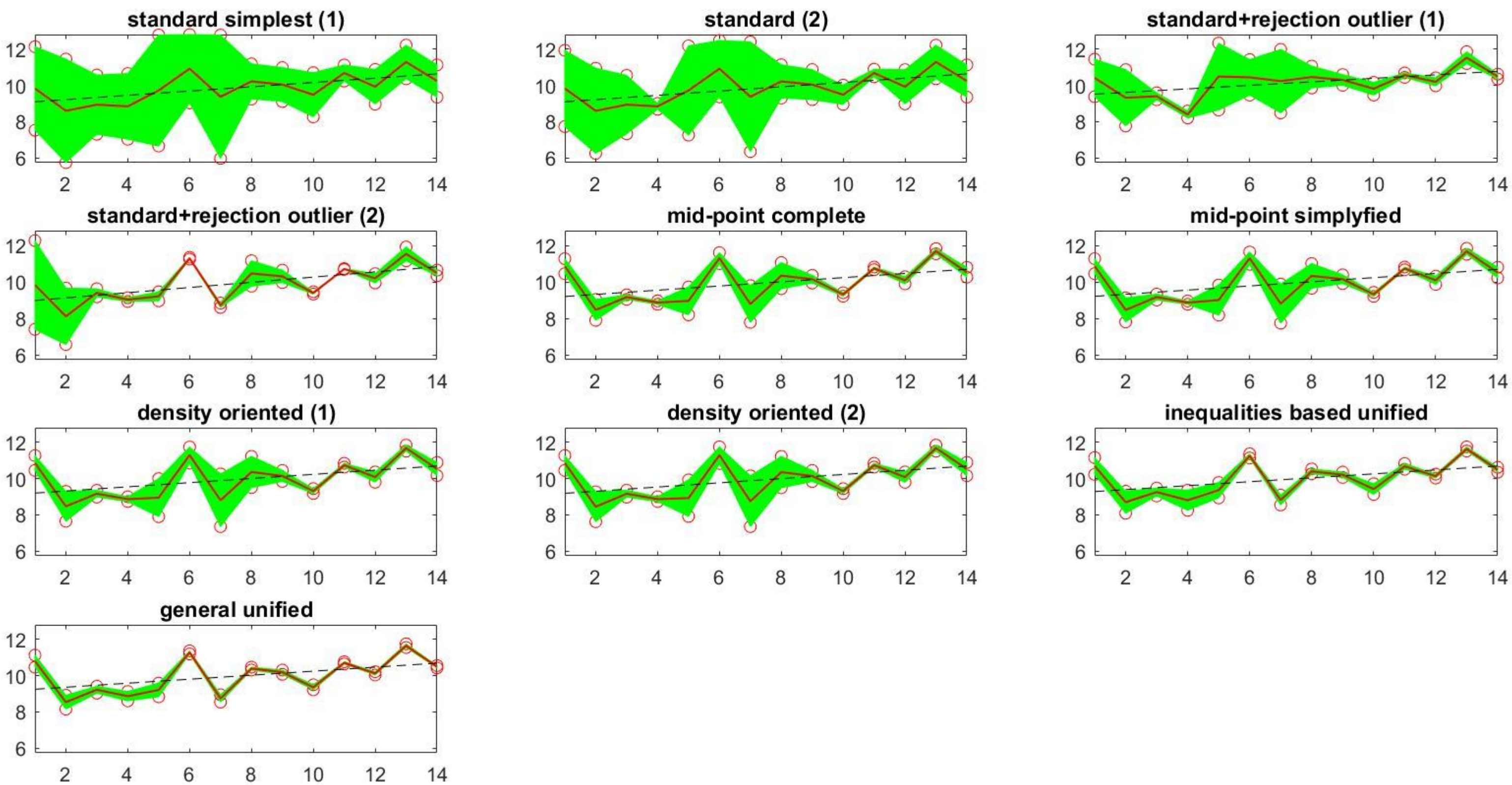
TRUE MEASUREMENTS - EXTENSOMETER



TYPICAL DATA REGISTRATION

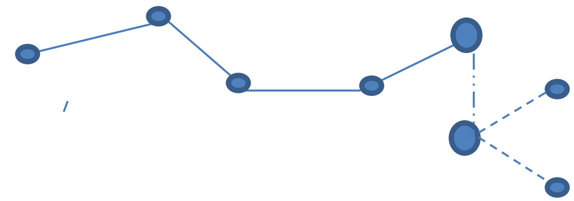


LEFT FLEXOR – TRUE MEASUREMENTS



APPLICATION OF STATISTICAL ANALYSIS RESULTS

1. Verification of results repeatability
2. Investigation of single patient
3. Interpretation of analysis results

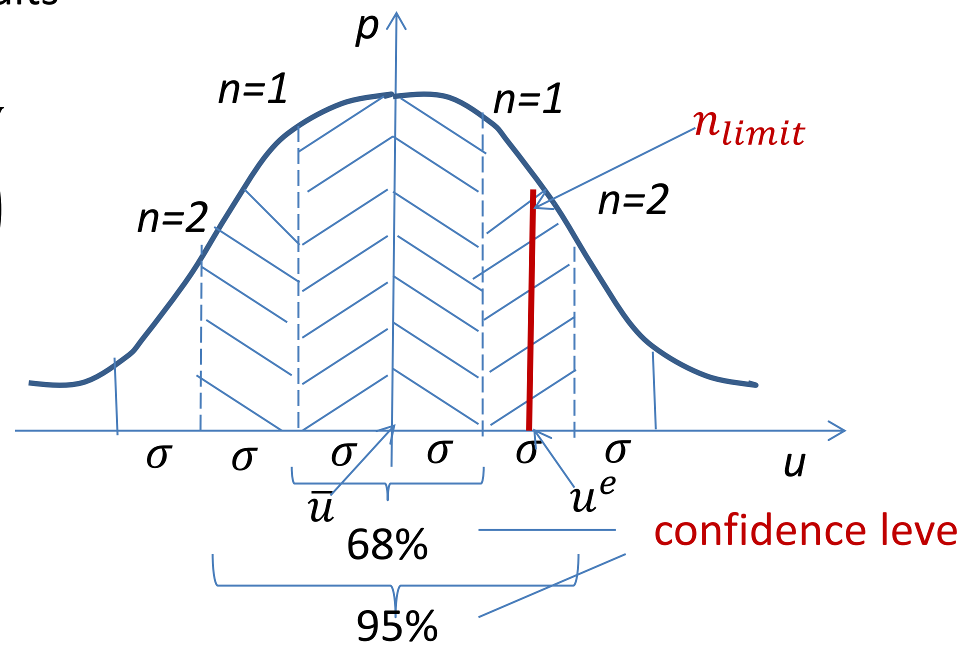


Gaussian probability density

$$p = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(u-\bar{u})^2}{2\sigma^2}\right)$$

confidence interval

$$\bar{u} - n\sigma \ \le \ u \ \le \ \bar{u} + n\sigma$$



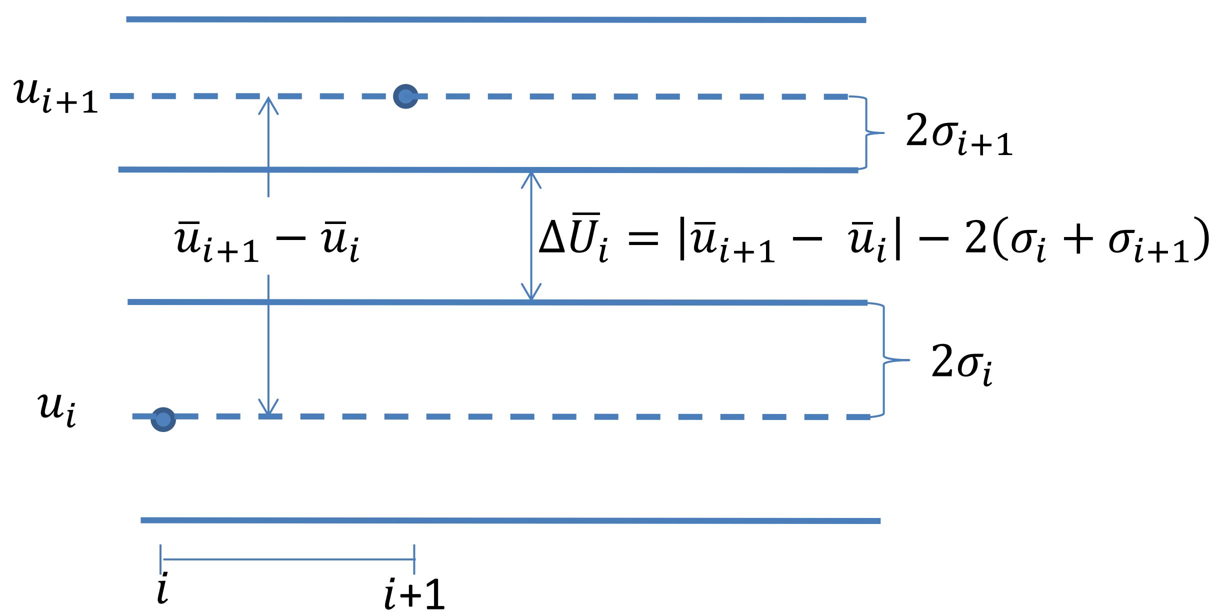
assume *n* (mostly *n* = 2)

answer quations

- does measured data $u^e \in [\bar{u} - n\sigma, \bar{u} + n\sigma]$?
- which is confidence level limit for u^e ?

$$\left.\begin{array}{l} u^e > 0 \ \rightarrow \ \bar{u} + n_{limit}\sigma \ = u^e \\ u^e < 0 \ \rightarrow \ \bar{u} - n_{limit}\sigma \ = u^e \end{array}\right\} \Rightarrow \ n_{limit}$$

TRUE AND STATISTICAL ERROR ZONES

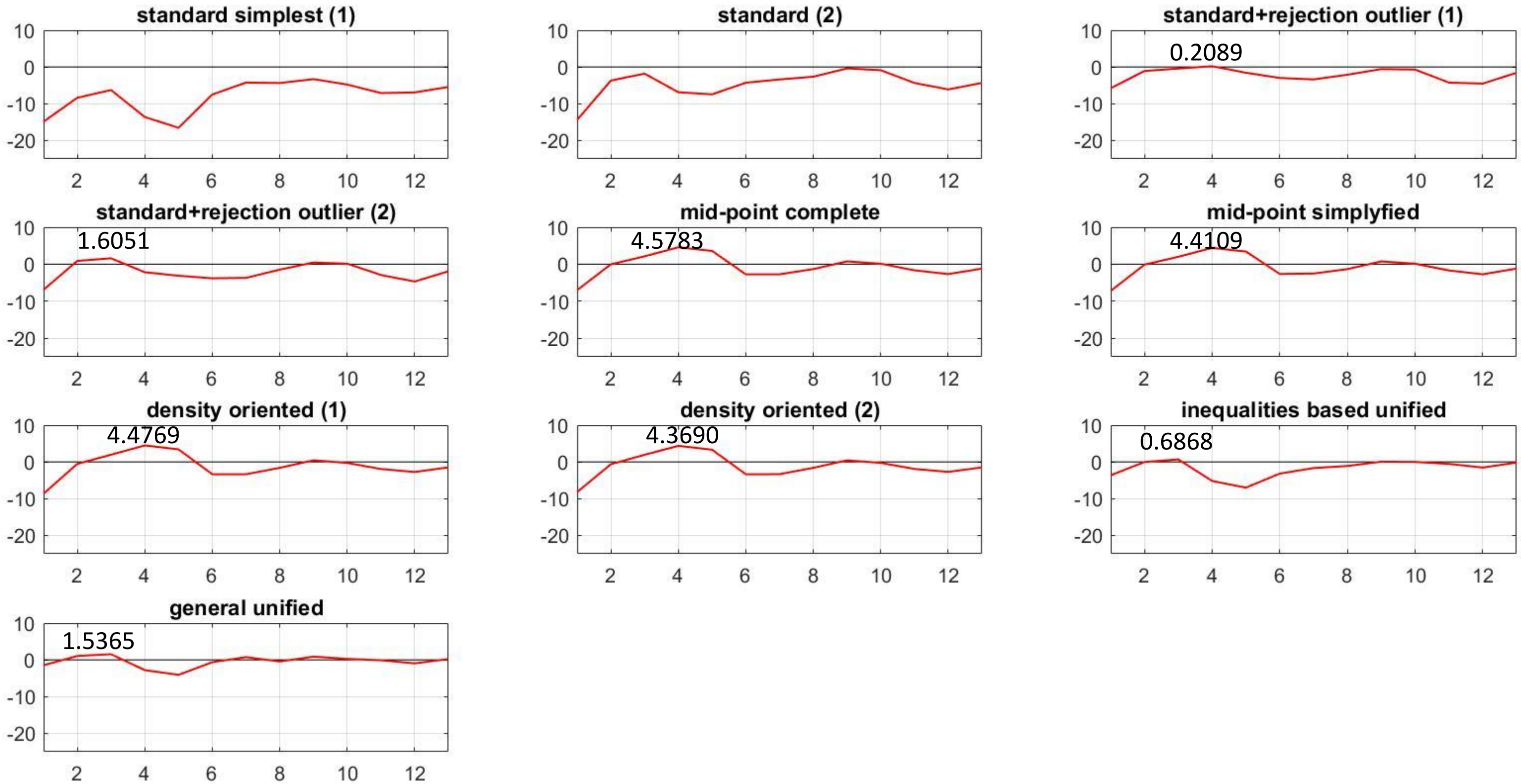


$\Delta\bar{U}_i > 0$ REAL CHANGES ZONE

$\Delta\bar{U}_i = 0$ LIMIT CASE

$\Delta\bar{U}_i < 0$ STATISTICAL CHANGES ZONE

FLEXOR P – STATISTICAL ERROR OR TRUE EFFECT ?



bmax = -3.3065 -0.3986 0.2089 1.6051 4.5783 4.4109 4.4769 4.3690 0.6868 1.5365

NEW SOLUTION APPROACH

MAIN CONCEPT   SUPPORT TOO LITTLE, POOR QUALITY DATA

BY RELEVANT NUMEROUS RANDOM PSEUDO DATA

RANDOM PSEUDO VECTORS DERIVATION

- START FROM THE „TRUE”   4   MEASURED DATA CHARACTERISTICS (u_{tm}, σ_{tm})
- GENERATE RANDOM VECTOR ${}_k\mathbf{u}_{rnd} = \{{}_k u_i\},\ i = 1,2,\ldots,k = 4$
- SPLIT VECTOR ${}_k\mathbf{u}$ INTO
  - EXPECTED VALUE PART $\bar{u}_{rnd}\mathbf{I}$
  - STANDARD DEVIATION PART σ_{rnd}
- INTRODUCE REDUCED RANDOM VECTOR ASSUMING $\boxed{\bar{u}_{red} = \bar{u}_{rnd}}$ $\Rightarrow$

$$\mathbf{u}_{red} = \bar{u}_{rnd}\mathbf{I} + \beta\,(\mathbf{u}_{rnd} - \bar{u}_{rnd}\mathbf{I}) \Longrightarrow \sigma_{red} = \beta\sigma_{\ rnd}$$

DEFINE β ASSUMING SCALING RANDOM DATA TO „TRUE” MEASUREMENTS

$$\boxed{\left[\frac{\sigma_{red}}{\bar{u}_{red}} = \frac{\sigma_{rnd}}{\bar{u}_{rnd}}\beta = \frac{\sigma_{tm}}{\bar{u}_{tm}}\right]} \implies \beta = \frac{\sigma_{tm}}{\bar{u}_{tm}}\left(\frac{\sigma_{rnd}}{\bar{u}_{rnd}}\right)^{-1}$$

INTRODUCE SIMULATED PSEUDO MEASUREMENTS VECTOR

$$\boxed{\mathbf{u}_{sim} = \frac{\bar{u}_{tm}}{\bar{u}_{rnd}}\mathbf{u}_{red}}$$

FINALLY RECEIVING

$$\boxed{\mathbf{u}_{sim} = \bar{u}_{tm} + \frac{\sigma_{tm}}{\sigma_{rnd}}\left(\mathbf{u}_{rnd} - \bar{u}_{rnd}\mathbf{I}\right)}$$

out of n

PRELIMINARY TEST:

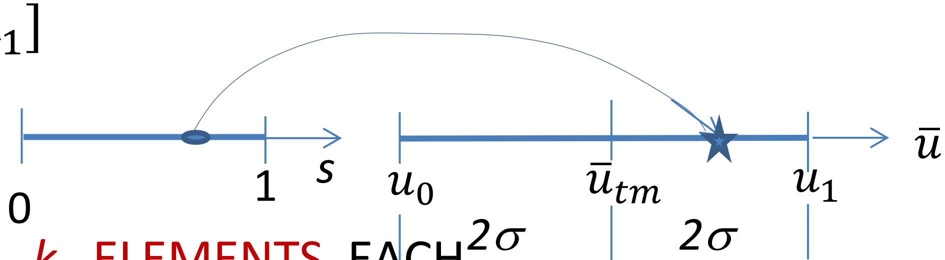
HOW ANALYSIS RESULTS DEPEND ON SUBSET *k* SIZE
(NUMBER OF TRUE MEASUREMENTS AVAILABLE)

- FIND $\bar{u}_{tm}, \sigma_{tm}$ FOR TRUE MEASUREMENTS $\{_k u_i\},\ i = 1,2,\dots,k = 4$
- USE *n* TIMES RANDOM *s* GENERATOR AND MAPPING

  $s \epsilon [0,1]$ ONTO INTERVAL $[u_0, u_1]$

  $u_1 = \bar{u}_{tm} + 2\sigma_{tm}$

  $u_0 = \bar{u}_{tm} - 2\sigma_{tm}$

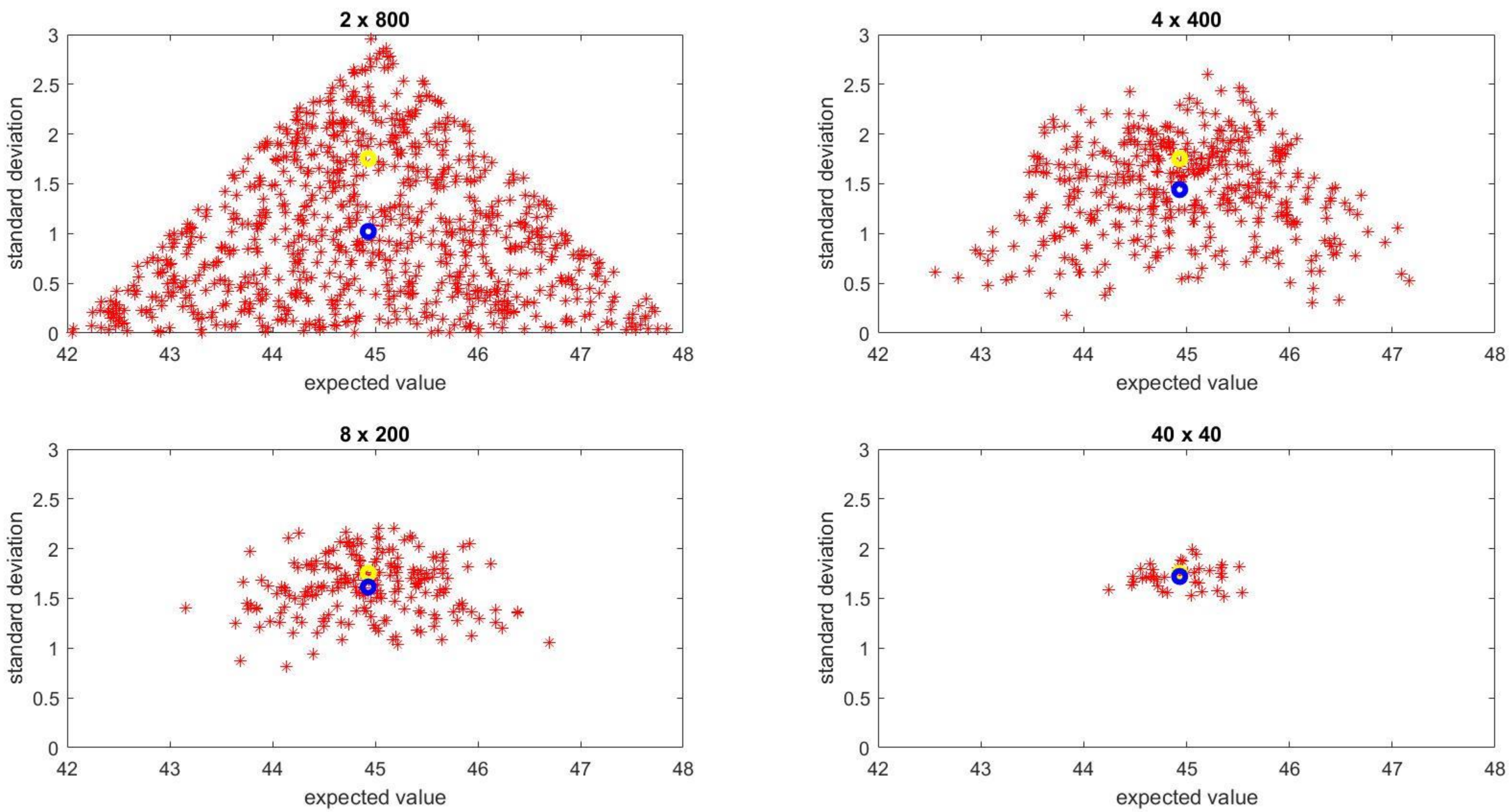


- FORM $\frac{n}{k}$ SUBSETS CONSISTING OF *k* ELEMENTS EACH
- FIND EXPECTED VALUE $\bar{u}$ AND STANDARD DEVIATION σ FOR EACH SUBSET
- GENERATE CLOUD OF $\frac{n}{k}$ POINTS $(\bar{u}, \sigma)_i$, $i = 1,2,\dots,\frac{n}{k}$

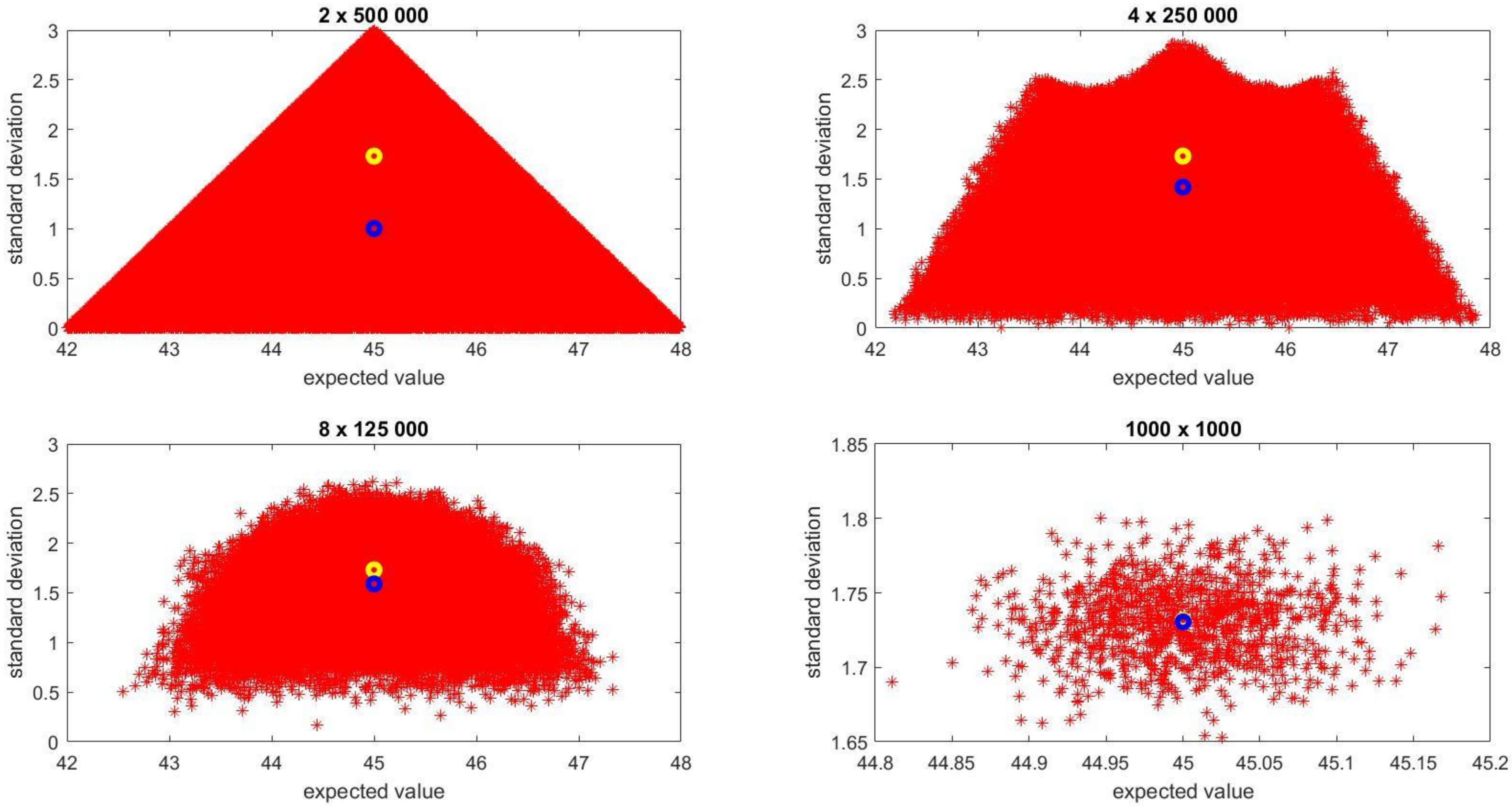
  AND THEIR MEAN VALUES (CLOUD CENTER OF GRAVITY: $\bar{u}_{AV}, \sigma_{AV}$)
- ASSUME *n*=1000000 AND *k*=2,4,5,8,10,20,25,40,50,80,100,125,200,250,400,500, 800,1000

  FIND RELATION $\sigma_{AV} = \sigma_{AV}(k)$

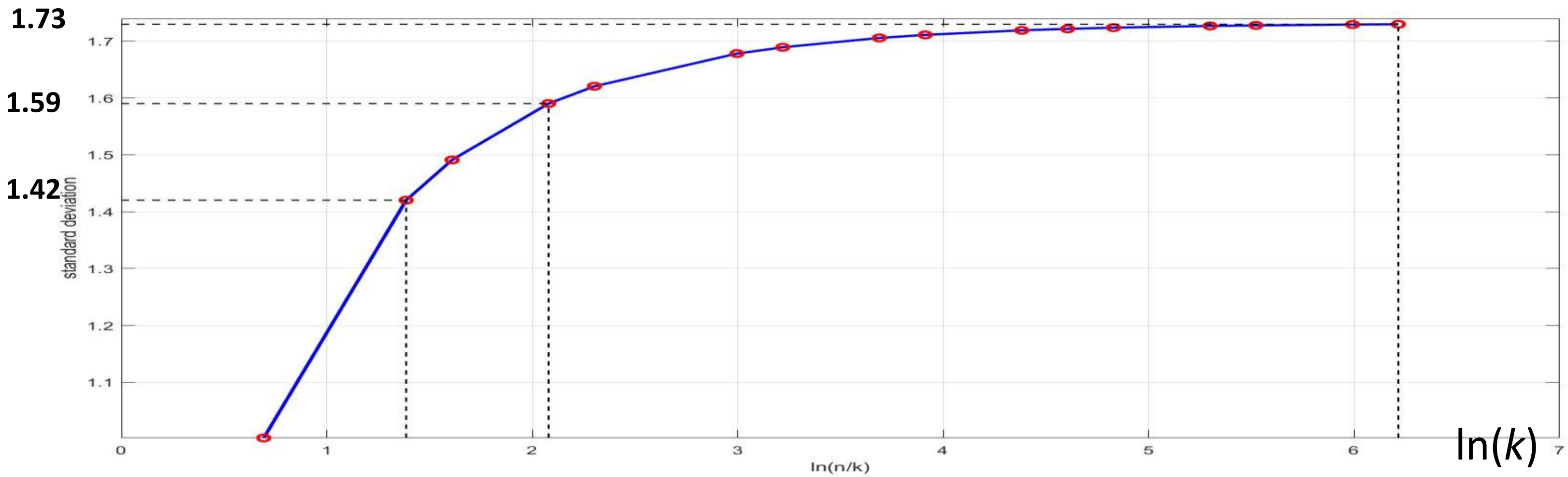
n=1600


2 x 800
4 x 400
8 x 200
40 x 40
standard deviation
expected value
0
0.5
1
1.5
2
2.5
3
42
43
44
45
46
47
48

n=1 000 000


2 x 500 000
4 x 250 000
8 x 125 000
1000 x 1000
standard deviation
expected value
0
0.5
1
1.5
2
2.5
3
42
43
44
45
46
47
48
1.65
1.7
1.75
1.8
1.85
44.8
44.85
44.9
44.95
45.05
45.1
45.15
45.2

LOCATION OF CLOUD GRAVITY CENTER - STANDARD DEVIATION $\sigma_{AV}(k)$



CONSTANT FACTOR (1 000 000) $\quad {}_4\mu = \dfrac{\sigma_{AV}(250000)}{\sigma_{AV}(4)} = \dfrac{1.73}{1.42} = 1.22$

RESULTS INTERPRETATION

$${}_k\mu = \frac{\text{NUMEROUS DATA } \ \sigma_{AV}\left(\frac{n}{k}\right)}{\text{TO LITTLE DATA } \ \sigma_{AV}\,(k)} = \quad \longleftrightarrow$$

CONSTANT VALUE RELATION

$\bar{u}_{tm}, \sigma_{tm}$ INDEPENDENT

- CLOUD SHAPE COMMENTS

GENERATION OF PSEUDO MEASUREMANTS

- START FROM EXPECTED VALUE $\bar{u}_{tm}$ AND STANDARD DEVIATION σ_{tm}

  FOUND FROM TRUE *k* (*k*=4) MEASUREMENTS

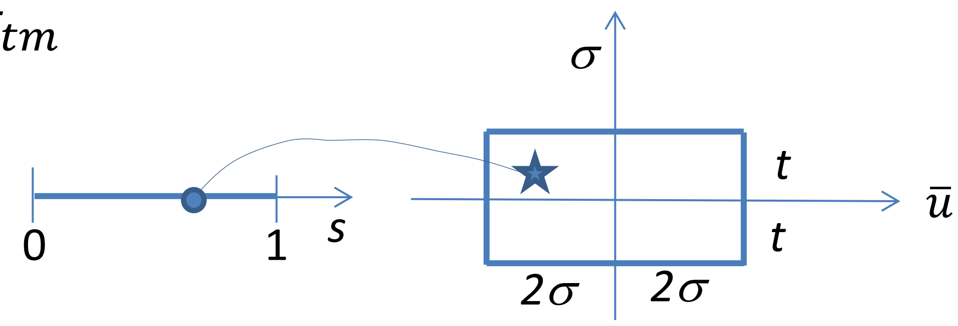
- SELECT 100 SIMULATED PSEUDO VECTORS $\mathbf{u}_{sim}$ CHOSEN FROM E.G. 200 ONES

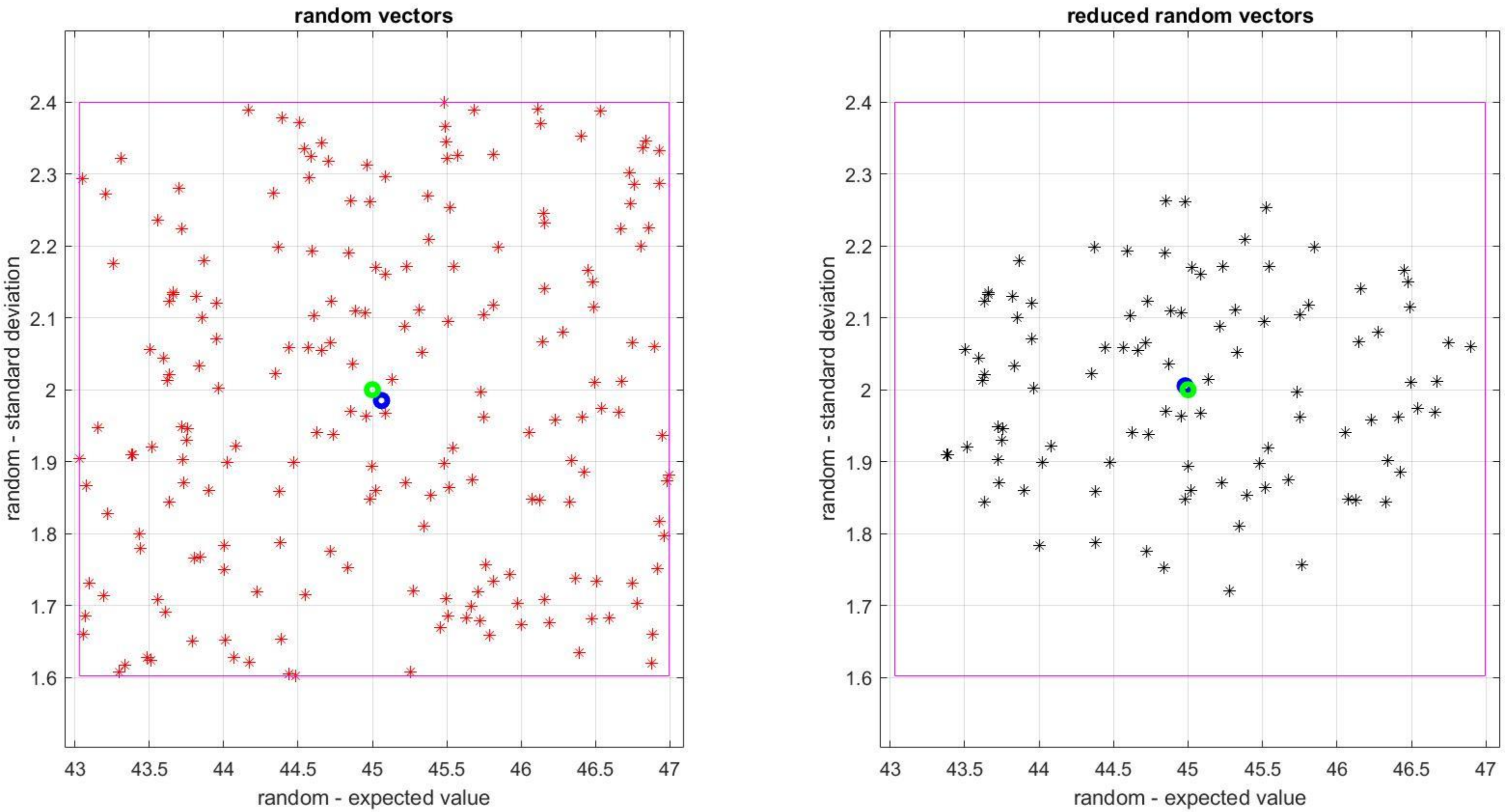
  OBTAINED BY USING RANDOM NUMBERS GENERATION $s \epsilon [0,1]$

  TRANSFORMED ONTO CONFIDENCE INTERVAL *s*

$$\bar{u} = 4\sigma_{tm}s + \bar{u}_{tm} - 2\,\sigma_{tm}$$

$$\sigma = 2ts + \sigma_{tm} - t$$




random vectors
reduced random vectors
random - standard deviation
random - expected value
2.4
2.3
2.2
2.1
2
1.9
1.8
1.7
1.6
43
43.5
44
44.5
45
45.5
46
46.5
47

- APPLY GAUSSIAN RANDOM NORMAL DISTRIBUTION FORMULAS

$$p(\bar{u}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\bar{u}-\bar{u}_{AV}}{2\sigma}\right)^2}$$

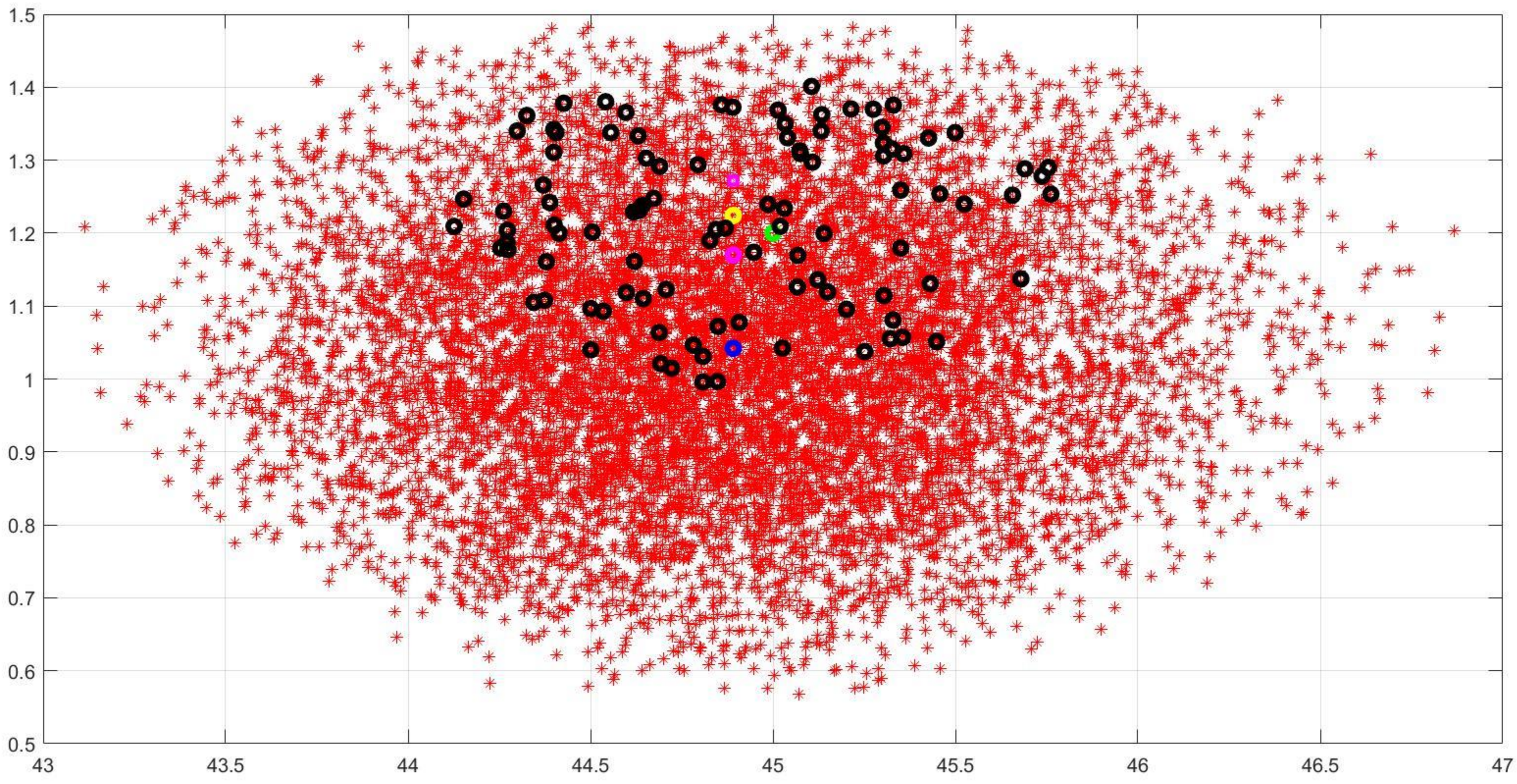
$$p(\sigma) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\sigma - \sigma_{AV}}{t}\right)^2}$$

$$p_i(\bar{u}, \sigma) = p(\bar{u}_i)p(\sigma_i)$$

- SELECT 100 POINTS WITH THE LARGEST PROBABILITY p_i
  IN THE SAME WAY FOR EACH OF SELECTED 100 PSEUDO VECTORS $\mathbf{u}_{sim}$ FIND PARAMETERS $(\bar{u}_{simAV}, \sigma_{simAV})$, GENERATE 4*100 NEW RANDOM DATA

- ASSEMBLE ALL n = 100*(4*100) = 40 000 RANDOM MEASUREMENTS AND APPLY, IN A SIMILAR WAY AS ABOVE, THE FINAL GLOBAL (STANDARD OR INNOVATIVE) STATISTICAL ANALYSIS IN ORDER TO OBTAIN POSSIBLY RELIABLE CONFIDENCE INTERVAL $[\bar{u} - \sigma,\ \bar{u} + \sigma]$

FINAL DATA SET AND RESULTS



FINAL ANALYSIS

USE FORMULAS

- LOCAL SUBSET BASED $\bar{u}_j = \frac{1}{k}\sum_{j=1}^{k} \bar{u}_j$, $\sigma_j{}^2 = \left[\frac{1}{k}\sum_{j=1}^{k}\left(u_j - \bar{u}_{AV}\right)^2\right]$

- CENTER OF GRAVITY OF ALL $\frac{n}{k}$ SUBSETS POINTS $\left(u_j, \sigma_j\right)$

$$\bar{u}_{AV} = \frac{k}{n}\sum_{j=1}^{n/k} \bar{u}_j \quad , \qquad \sigma_{AV} = \frac{k}{n}\sum_{j=1}^{n/k} \sigma_j \quad ,$$

- STANDARD DEVIATION FOR ALL $\frac{n}{k}$ VALUES OF σ_j

$$\Delta\sigma_{AV}{}^2 = \frac{k}{n}\sum_{j=1}^{n/k}\left(\sigma_j - \sigma_{AV\,j}\right)^2$$

- GLOBAL STANDARD DEVIATION FORMULA FOR n MEASUREMENTS

$$\sigma_g{}^2 = \frac{1}{n}\sum_{j=1}^{n}\left(\bar{u}_j - \bar{u}_{AV}\right)^2$$

FINAL RESULTS

FIND FINAL CONFIDENCE INTERVALS USING VARIOUS CONCEPTS

LOCAL (FOR TRUE MEASUREMENTS)

$$u \in [\bar{u}_{tm} - \sigma_{tm}, \bar{u}_{tm} + \sigma_{tm}] = [45 - 1.2, 45 + 1.2] = [43.80, 46.20]$$

GLOBAL

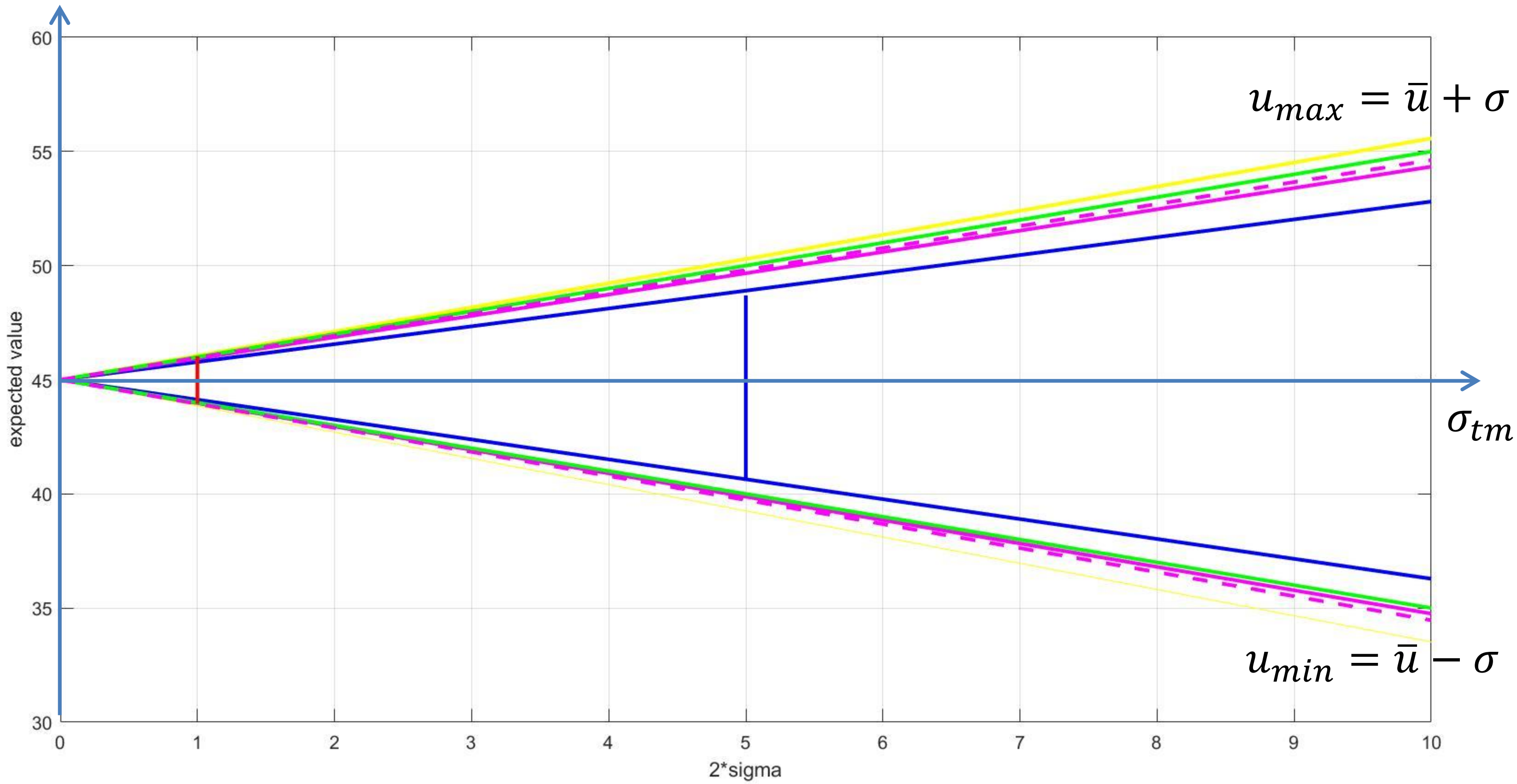
$$u \in [\bar{u}_{AV} - \sigma_{AV}, \bar{u}_{AV} + \sigma_{AV}] = \quad [44.01, 46.06]$$

$$u \in [\bar{u}_{AV} - \sigma_{AV} - \Delta\sigma_{AV}, \bar{u}_{AV} + \sigma_{AV} + \Delta\sigma_{AV}] = \quad [43.88, 46.19]$$

$$u \in [\bar{u}_{AV} - \mu\sigma_{AV}, \bar{u}_{AV} + \mu\sigma_{AV}] = \quad [43.78, 46.29]$$

$$u \in \left[\bar{u}_{AV} - \sigma_{g}, \bar{u}_{AV} + \sigma_{g}\right] = \quad [43.83, 46.24]$$

FINAL RESULTS



confidence interval

FINAL REMARKS

- RANDOM PSEUDO MEASUREMENTS SUPPORT FOR TOO LITTLE AND POOR QUALITY DATA WAS CONSIDERED
- SEVERAL VARIANTS OF THE APPROACH WERE INVESTIGATED AND DISCUSSED
- ALL FINAL RESULTS OF THE METHOD CONSIDERED YIELD PRETTY CLOSE RESULTS FOR BOTH LOCAL AND GLOBAL CONFIDENCE INTERVALS
- LET US LISTEN AGAIN WHAT PESSIMISTS AND OPTIMISTS COULD SAY NOW

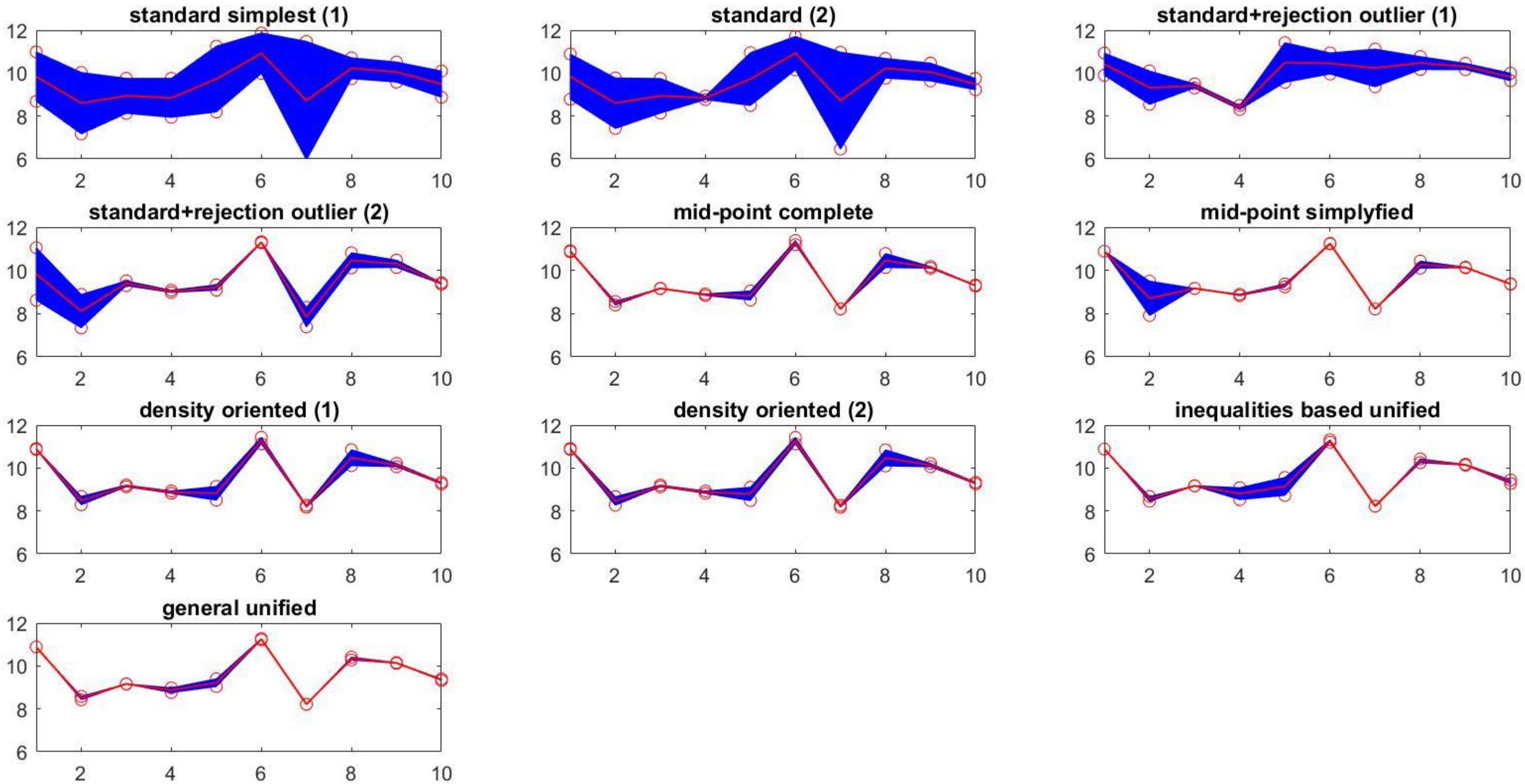
  PESSIMISTS: WE DEPARTED FROM THE $\bar{u}_{tm}, \sigma_{tm}$ DATA AND DUE TO STATISTICAL ANALYSIS WE RETURNED AGAIN TO THE SAME SPOT, NOTHING WAS GAINED THEN

  OPTIMISTS: RESULTS OF ALL PROPOSED WAYS OF INNOVATIVE RANDOM DATA ANALYSIS ARE CLOSE ENOUGH – THEREFORE, THIS FACT CONFIRMS THE APPROACH
- WHERE IS THE TRUTH THEN?

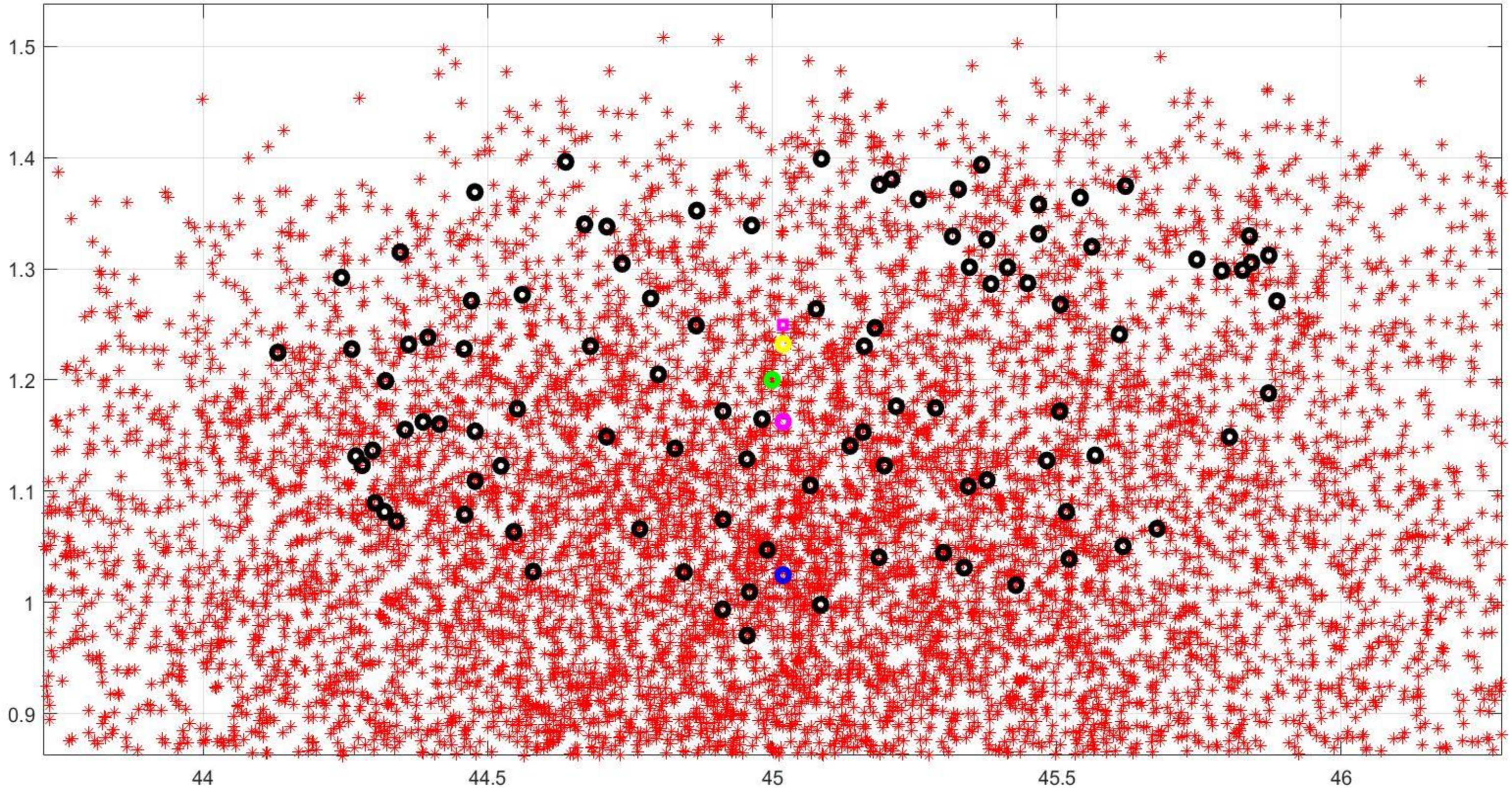
  LET US KEEP IT OPEN UNTIL A VERIFICATION DONE BY ANALYSIS OF SUFFICIENTLY LARGE AMOUNT OF TRUE EXPERIMENTAL DATA COULD BE AVAILABLE AND TESTED

THANK YOU FOR ATTENTION

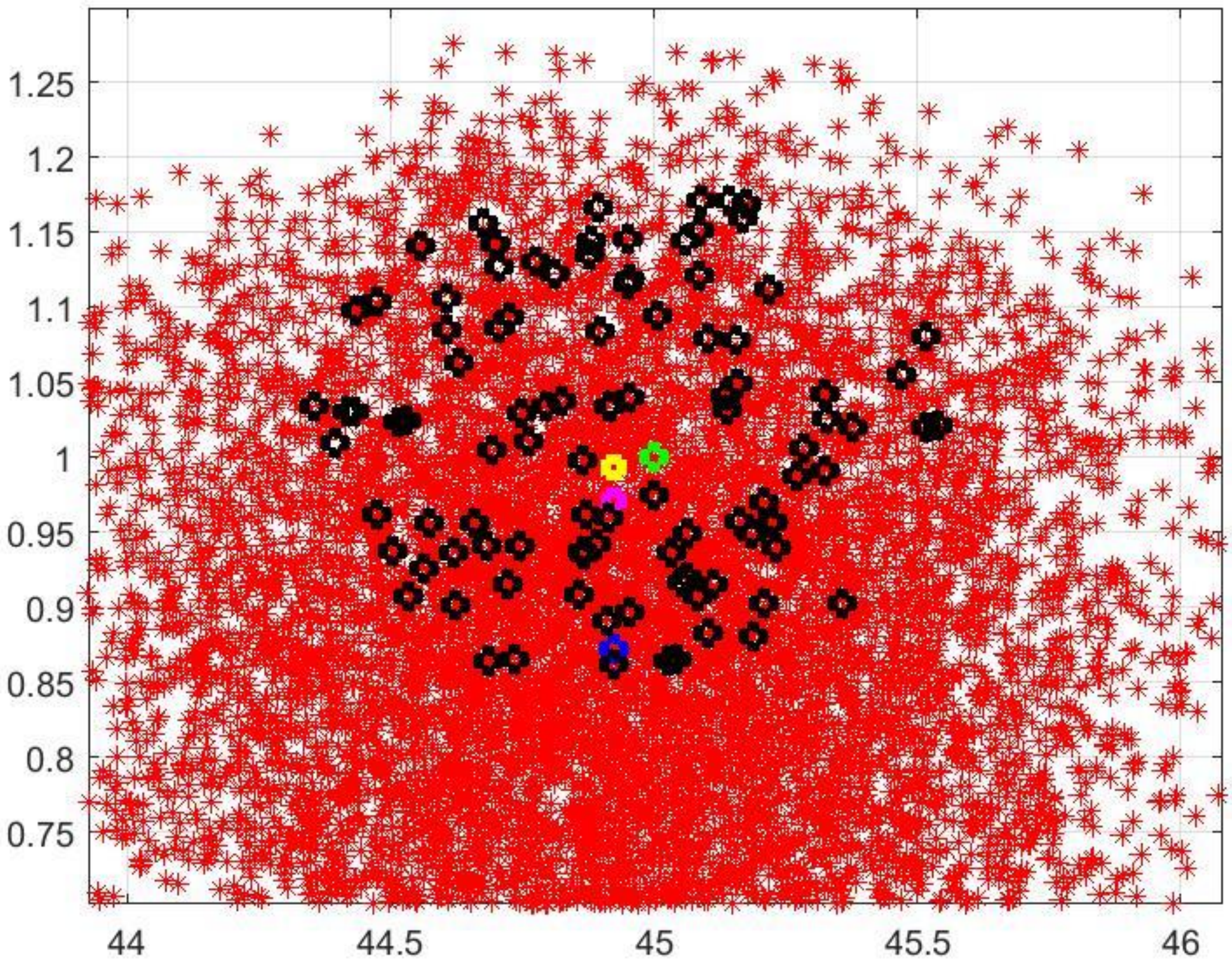
LEFT FLEXOR, k=2

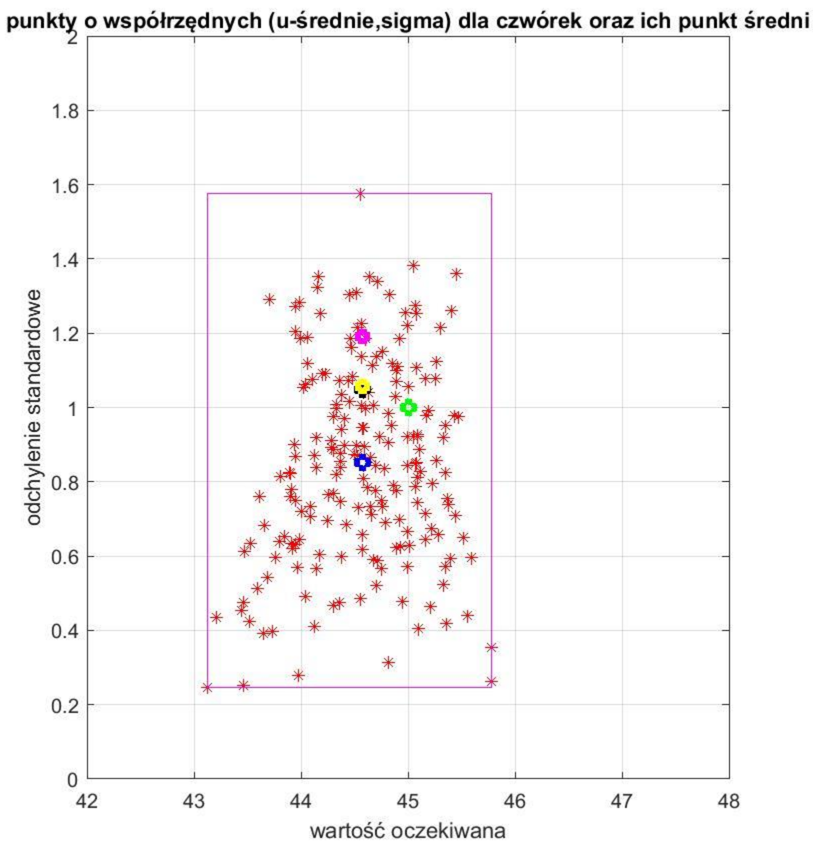


FINAL DATA SET AND RESULTS

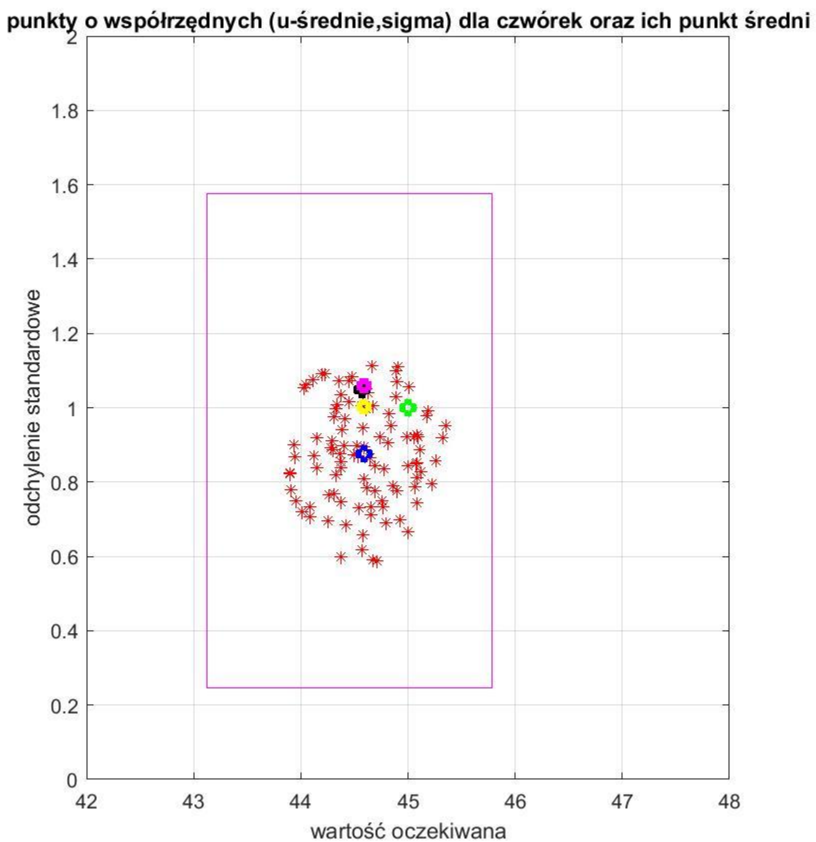


FINAL DATA SET AND RESULTS





punkty o współrzędnych (u-średnie,sigma) dla czwórek oraz ich punkt średni
odchylenie standardowe
wartość oczekiwana
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
42
43
44
45
46
47
48



punkty o współrzędnych (u-średnie,sigma) dla czwórek oraz ich punkt średni
odchylenie standardowe
wartość oczekiwana
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
42
43
44
45
46
47
48