A picture containing text, clipart

Description automatically generatedLogo, company name

Description automatically generated**Универзитет у Нишу**

**Електронски факултет**

**Катедра за рачунарство**

**Технике и методе анализе података**

**Детекција тачака промене (*Change point detection*)**

|  |  |  |
| --- | --- | --- |
| **Ментор**: |  | **Кандидат**: |
| Проф. др. Сузана Стојковић |  | Стеван Грујић 1493 |

**Ниш, 2023.**

Садржај

[1 Увод 3](#_Toc144134679)

[1.1 Основе детекција тачака промене 3](#_Toc144134680)

[1.2 *Offline* и *Online Change Point detection* 3](#_Toc144134681)

[1.3 Библиотека *Ruptures* 4](#_Toc144134682)

[2 Практични део пројекта 4](#_Toc144134683)

[2.1 Скуп података 5](#_Toc144134684)

[2.2 Коришћени алгоритми детекције тачака промене 6](#_Toc144134685)

[2.2.1 *PELT* алгоритам 6](#_Toc144134686)

[2.2.2 *BinSeg* алгоритам 8](#_Toc144134687)

[2.2.3 *BottomUp* алгоритам 9](#_Toc144134688)

[2.3 Ток анализе података 10](#_Toc144134689)

[2.3.1 Визуелизација података 10](#_Toc144134690)

[2.3.2 Претпроцесирање података 13](#_Toc144134691)

[2.3.3 *Change point detection (Univariate)* 15](#_Toc144134692)

[2.3.4 *Change point detection (Multivariate)* 18](#_Toc144134693)

[3 Закључак 21](#_Toc144134694)

[4 Литература 22](#_Toc144134695)

# Увод

## Основе детекција тачака промене

У савременом друштву, константне промене постале су неизбежна карактеристика динамичних система и процеса у различитим сферама људске делатности. С обзиром на ову динамичност, способност препознавања, праћења и адаптације на промене постала је од суштинског значаја за одржавање ефикасности, поузданости и успешности различитих система. У том контексту, анализа тачака промене, позната као "*Change point detection*" (*CPD*), издваја се као кључна техника која омогућава идентификацију критичних тренутака када се статистичка својства посматраних података значајно мењају.

Основни задатак *CPD*-а је препознати промене у статистичким карактеристикама података, као што су средња вредност, варијанса, корелација или спектрална густина процеса. Конкретно, ова техника се бави детектовањем промена у расподелама вероватноћа података, идентификујући тренутке када се те расподеле значајно разликују. То може укључивати откривање промена у средњој вредности, варијанси, корелацији или спектралној густини процеса. Осим тога, *CPD* такође обухвата детекцију аномалија или изузетног понашања у подацима, што се назива и "*anomaly detection*".

*CPD* налази примену у широком спектру дисциплина, укључујући индустрију, економију, медицину, телекомуникације и многе друге области. Ова техника омогућава аналитичарима и истраживачима да дубље разумеју процесе, идентификују аномалије, оптимизују перформансе и самим тим доносе боље одлуке. Кључно је истаћи да правовремено препознавање промена може спречити губитке, побољшати ефикасност и допринети конкурентсној предности.

## *Offline* и *Online Change Point detection*

Детекција тачака промене представља кључни задатак у анализи временских серија података, јер омогућава идентификацију тренутака када се понашање система значајно мења. "*Online*" алгоритми су усмерени на брзу детекцију и одговарање на промене у стварном времену. Они су посебно корисни у ситуацијама где је потребно реаговати моментално, као што је откривање аномалија у индустријским процесима или праћење сензора у *IoT* (*Internet of Things*) окружењима. Пример "*online*" алгоритма је рекурентна метода помака (*Recursive Moving Average*) која прати средњу вредност сигнала током времена и детектује промене када тренутна вредност одступа од очекиване вредности за одређени праг.

Са друге стране, "*offline*" алгоритми имају за циљ пружити дубљу анализу и прецизнију детекцију промена након што је цео сигнал прикупљен. Ови алгоритми омогућавају свеобухватно разумевање промена у дужем временском периоду и могу бити корисни за истраживање узрока и последица. На пример, алгоритам базиран на сегментацији сигнала као што је "*Change point detection via Kuliback-Leibler segmentation*" разбија сигнал на сегменте са сличним карактеристикама и идентификује тачке промене међу сегментима на основу статистичких метода.

Иако се "*offline*" алгоритми генерално користе након што су подаци прикупљени, постоји и средњи приступ који се назива "*online batch*" детекција. Овај приступ комбинује елементе оба света, омогућавајући обраду података у "*offline*" маниру, али у оквиру временског прозора. На пример, алгоритам који користи клизни прозор ("*sliding window*") за анализу последњег временског интервала може се сматрати хибридним, јер омогућава реалном времену анализу са високим степеном тачности.

У суштини, избор између *online* и *offline* приступа зависи од специфичних захтева апликације, брзине реакције која је потребна и дубине анализе која се очекује. Треба напоменути да се овај рад искључиво бави *offline* приступом.

## Библиотека *Ruptures*

Осим поменутих традиционалних приступа детекцији тачака промене, у новије време, истраживачи се све више ослањају на различите алате и библиотеке које олакшавају ову анализу. Једна од таквих библиотека је "*ruptures*", која је постала вероватно и најпопуларнији избор у многим истраживачким радовима. *Ruptures* библиотека пружа широк спектар алгоритама за детекцију тачака промене у временским серијама, омогућавајући истраживачима да брзо и ефикасно анализирају своје податке.

*Ruptures* нуди подршку за различите приступе, укључујући "*offline*" сегментацију и анализу временских серија. Библиотека укључује имплементацију алгоритама као што су *PELT*, *BinSeg*, *и KernelCPM*, који омогућавају анализу промена у дистрибуцијама података, претрагу по растојању између сегмената и друге методе. Осим тога, *ruptures* омогућава и лаку употребу класичних статистичких метода као што су *CUSUM* (*Cumulative Sum*) и *Hinkley* тест за детекцију промена.

Кроз употребу *ruptures* библиотеке, истраживачи су у могућности да примене различите алгоритме и приступе у детекцији тачака промене на својим подацима, без потребе за имплементацијом алгоритама из почетка. Ова библиотека је допринела унапређењу ефикасности и тачности анализе промена у временским серијама, чинећи процес истраживања и откривања информација лакшим и бржим.

# Практични део пројекта

У овом поглављу биће речи о скупу података над којим је вршена анализа, као и о алгоритмима који су примењени са циљем детекције тачака промене у одговарајућим временским серијама.

## Скуп података

Скуп података који је коришћен у овом истраживању је назван "*Enviromental data*" и састоји се од временских серија које су забележене помоћу сензора смештених унутар одређене просторије у фабрици. Ови подаци су аквизирани континуирано током четвородневног периода са високом фреквенцијом одабирања од једне секунде. Скуп података омогућава дубоку анализу фактора који утичу на квалитет околине унутар просторије.

У наставку следи опис појединачних атрибута поменутог скупа података:

**Временска Ознака (*timestamp*)**: Сваки забележени податак је праћен временском ознаком која прецизно означава тренутак када је податак снимљен. Овај атрибут омогућава синхронизацију са другим релевантним догађајима и променама у окружењу.

**Звук (*sound*)**: Вредност звука, изражена у децибелима, рефлектује ниво буке у просторији током датог тренутка. Прецизно праћење овог атрибута може указати на периодичне или акутне измене у радном окружењу.

**Атмосферски притисак (*pressure*)**: Ова вредност, изражена у хектопаскалима, представља меру спољашњег атмосферског притиска у датом тренутку. Флуктуације овог параметра могу бити последица метеоролошких промена које утичу на унутрашње услове.

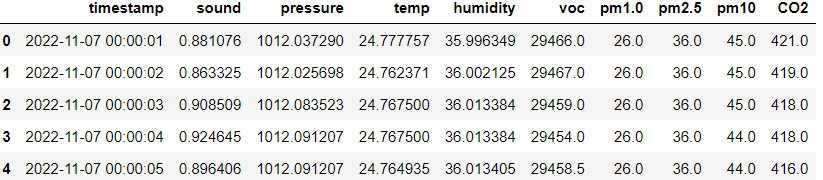
**Температура ваздуха (*temp*)**: Вредност температуре ваздуха, изражена у степенима Целзијуса, пружа увид у температурне варијације у окружењу. Овај податак је од суштинског значаја за оцену удобности и адекватности услова рада.

**Релативна влажност ваздуха (*humidity*)**: Овај атрибут, изражен у процентима, означава степен засићености ваздуха воденом паром. Промене релативне влажности могу утицати на квалитет ваздуха и удобност радног простора.

**Вредност органских једињења (*voc*)**: Мерење органских једињења у ваздуху, изражено у деловима по милиону (ппм), указује на присуство потенцијално штетних супстанци. Овај атрибут је од значаја за оцену квалитета ваздуха у смислу контаминације.

**Број честица у ваздуху (*pm*1.0, *pm*2.5, *pm*10)**: Три различите величине честица, мерене у броју честица по кубном метру, пружају увид у присуство суспендованих честица у ваздуху. Ово је посебно важно за праћење нивоа загађења ваздуха.

**Угљен-диоксид (*CO2*)**: Концентрација угљен-диоксида у ваздуху, изражена у ппм, указује на ниво присуства издаха и вентилацију просторије. Ова вредност је индикатор ефикасности вентилационог система.



**Слика 2.1.1** „*Enviromental data*“

Поменути атрибути заједно чине свеобухватан скуп информација о стању околине унутар анализиране фабричке просторије током четвородневног периода. Комбиновањем ових података омогућено је дубинско разумевање динамике фактора који утичу на квалитет окружења, што представља основу за даљу анализу и закључивање у наредним поглављима истраживања.

## Коришћени алгоритми детекције тачака промене

Детекција тачака промене представља сложен проблем у анализи временских серија. Алгоритми за детекцију тачака промене имају за циљ аутоматски идентификовати ове тренутке према статистичким разликама између сегмената серије. Ови алгоритми омогућавају објективан приступ идентификацији значајних промена и имају широку примену у различитим областима као што су економија, медицина и инжењеринг.

У склопу *ruptures* библиотеке постоје две врсте алгоритама. Прва врста се односи на алгоритме који захтевају постојање истинитих вредности тачака промене унапред (*ground truth*). Ови алгоритми се често користе у надгледаним сценаријима где имамо референтне тачке промене које се користе за евалуацију перформанси детекције.

Са друге стране, постоје алгоритми који не захтевају постојање референтних тачака промене и спадају под алгоритме намењене за ненадгледане сценарије.

У оквиру овог пројекта коришћена је друга група алгоритама (алгоритми за ненадгледане сценарије) с обзиром на то да овај скуп података није садржао референтне тачке промене (*ground truth*). Конкретни алгоритми који су коришћени су: ***PELT*** (*Pruned Exact Linear Time*),***BinSeg***(*Binary Segmentation*), ***BottomUp***(*Bottom-Up Segmentation*). У наставку ће наведени алгоритми бити објашњени детаљније.

### *PELT* алгоритам

*PELT* (*Pruned Exact Linear Time*) је алгоритам који се користи за откривање тачака промене у временским низовима или низовима података. Главна предност *PELT* алгоритма је брза и прецизна идентификација тачака промене уз ефикасно коришћење ресурса.

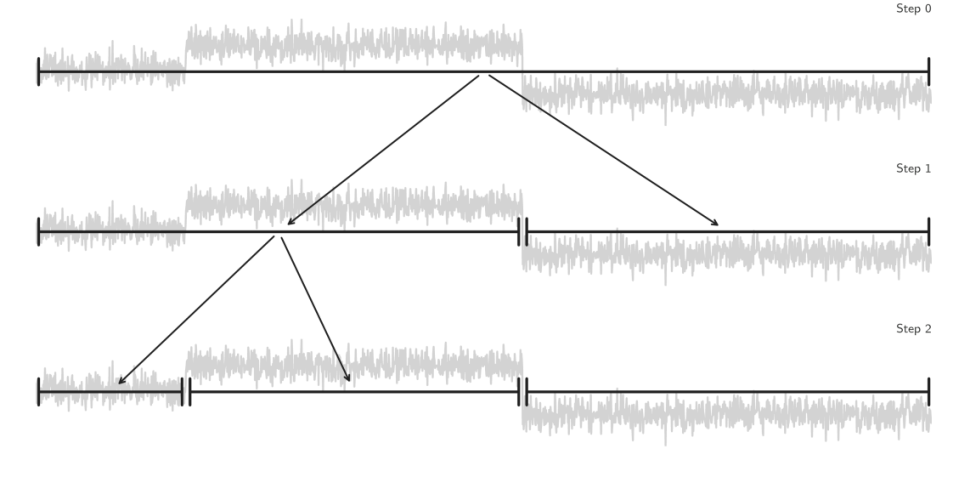
Ево како *PELT* алгоритам функционише:

1. **Сегментација на почетку**: Алгоритам почиње тако што сваки временски корак сматра засебним сегментом. То значи да на почетку сваки корак представља почетак нове сегментације.
2. **Рачунање кумулативних трошкова**: За сваки сегмент (комбинацију корака) алгоритам рачуна кумулативни трошак сегментације. Овај трошак мери колико добро одабрана сегментација одговара подацима. Трошак се често рачуна користећи одговарајуће метрике зависно од проблема (нпр. квадратна разлика за временске низове).
3. **Прилагођавање са казнама**: *PELT* додаје "казне" за сваку потенцијалну тачку промене. Ова казна представља цену која се плаћа за додавање нове тачке промене. Циљ је да се избалансира тачка промене која значајно смањује трошак сегментације са ценом коју плаћамо за увођење нове тачке промене.
4. **Селективно додавање тачака промене**: Алгоритам пролази кроз сваку тачку у низу и процењује да ли ће увођење те тачке као промене донети значајно смањење укупног трошка сегментације. Ако је тачка промене корисна у том смислу и ако корист надмашује казну, та тачка се додаје као нова тачка промене.
5. **Рекурзивна природа**: *PELT* алгоритам је рекурзиван - то значи да се итеративно примењује на све краће сегменте који потенцијално садрже тачке промене. То омогућава алгоритму да идентификује тачке промене унутар већих сегмената.
6. ***Pruning* (искључивање)**: Кључна карактеристика *PELT* алгоритма је "искључивање" непотребних сегмената. Када алгоритам пронађе тачку промене, она се користи као граница сегмента. То значи да се каснији делови сегмента не разматрају јер би увођење додатних тачака промене унутар тог сегмента резултирало лошијим укупним трошком.
7. **Глобална минимизација**: *PELT* алгоритам има за циљ да пронађе комбинацију тачака промене која минимизира укупни трошак сегментације, узимајући у обзир додате казне. Овај приступ осигурава да се открију значајне тачке промене без прекомерног сегментирања.

Све у свему, *PELT* алгоритам је софистицирани метод за откривање тачака промене у временским низовима. Његова ефикасност долази из комбинације прилагођеног приступа и рекурзивне примене, омогућавајући брзо и прецизно откривање промена у великим низовима података.

### *BinSeg* алгоритам

Детекција бинарних тачака промене (*Binary segmentation*) је ефикасан метод за брзу сегментацију сигнала. Ова техника се ослања на секвенцијални приступ: најпре се идентификује једна тачка промене у целокупном улазном сигналу. Након тога, сигнал се дели на два дела око ове детектоване тачке промене. Тада се исти поступак примењује на ова два нова сегмента, понављајући овај циклус све док се не постигне жељени ниво сегментације.



**Слика 2.2.2.1** „*BinSeg*“ илустрација

Принцип рада је прилично интуитиван. Можемо да замислимо сигнал као путовање кроз различите "режиме" или "фазе". Свака тачка промене означава прелазак из једног режима у други, као што је прелазак из раста у падање у временској серији цена акција. Бинарна сегментација покушава да пронађе ове тачке промене и идентификује границе између различитих фаза сигнала.

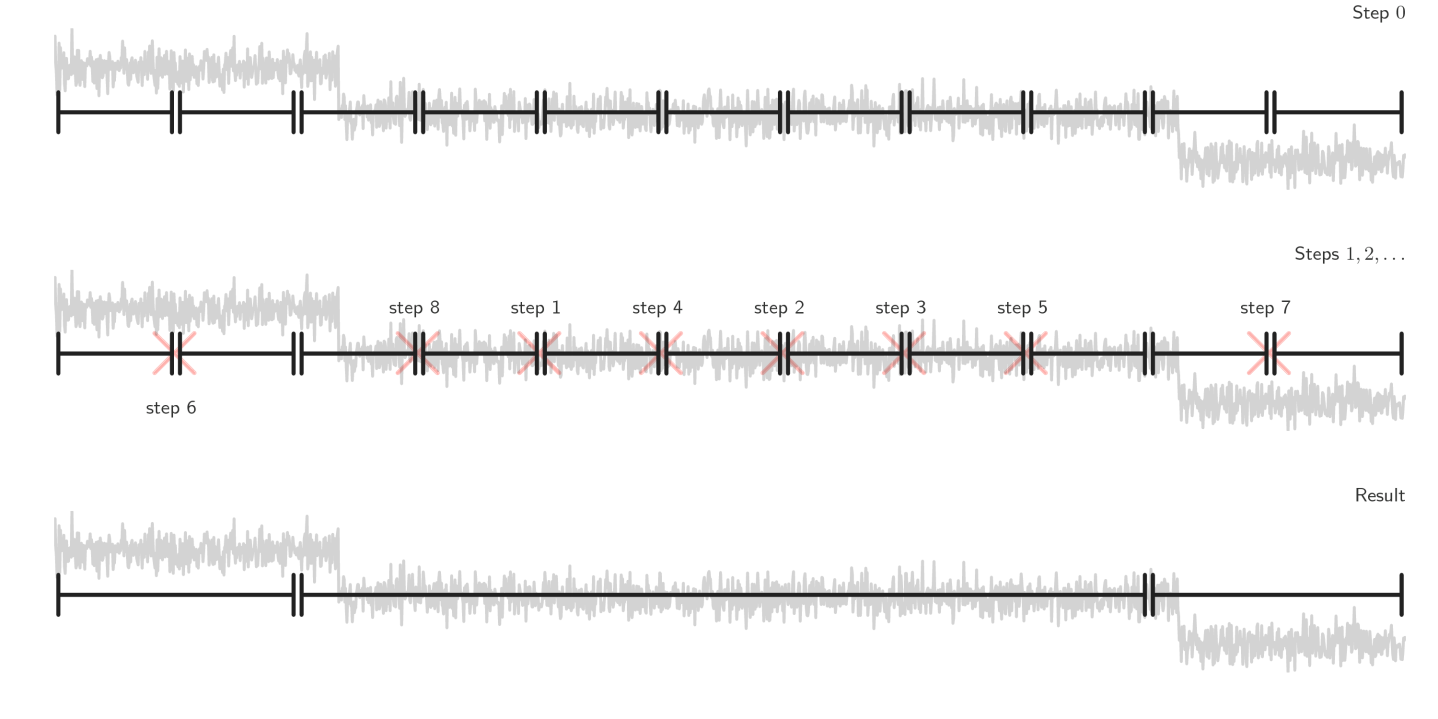
Ова метода има ниску сложеност, што значи да је изузетно ефикасна за примену на великим скуповима података. Сложеност операција расте логаритамски са бројем узорака (реда величине *O(n log n)*), што је кључно за брзе анализе у реалном времену. Осим тога, бинарна сегментација може проширити било коју постојећу методологију за детекцију појединачних тачака промене, омогућавајући откривање више тачака промене у сигналима.

Једна од значајних предности овог приступа је та што може радити без обзира на то да ли је број тачака промене унапред познат или не. То значи да је могуће применити ову технику и када немамо претходно знање о броју сегмената у сигналу.

Све у свему, детекција бинарних тачака промене је корисна техника за брзу и ефикасну сегментацију сигнала, омогућавајући нам да боље разумемо промене и прелазе унутар временских серија или других континуираних података.

### *BottomUp* алгоритам

Детекција тачака промене одоздо према горе представља секвенцијални приступ који обавља суштинску улогу у постизању ефикасне сегментације сигнала. Супротно од свог бинарног сегментационог еквивалента (*Binary Segmentation*), овај метод примењује поступак знатно обухватнијег карактера. Процес започиње са великим бројем тачака промене након чега постепено елиминише оне мање битне. У првој фази, сигнал се партиционише на низ сегмената, сви прецизно поравнати унутар стандардне мреже. Након тога, узастопни сегменти се поступно сједињују узимајући у обзир њихове заједничке карактеристике и сличности. Овај поступак приказан је на слици 2.2.3.1.



**Слика 2.2.3.1** „*Bottom-Up Segmentation*“ илустрација

Привлачност овог приступа проистиче из његове елегантне једноставности, карактерисане рачунарском сложеношћу реда *О(nlogn)*, где '*n*' означава број узорака. Осим ефикасности, овај метод има способност да прошири појединачне механизме детекције тачака промене како би обухватио вишеструке случајеве транзиције. Такође, слично бинарној сегментацији, лако се прилагођава различитим сценаријима и броју режима, било да су унапред познати или не.

На основу својих карактеристика, детекција тачака промене одоздо према горе се истиче као моћан алат за брзу сегментацију сигнала. Кроз пажљиво уклањање мање значајних тачака промене и способност прилагођења различитим методологијама, овај приступ се прецизно креће кроз сложену топографију анализе сигнала.

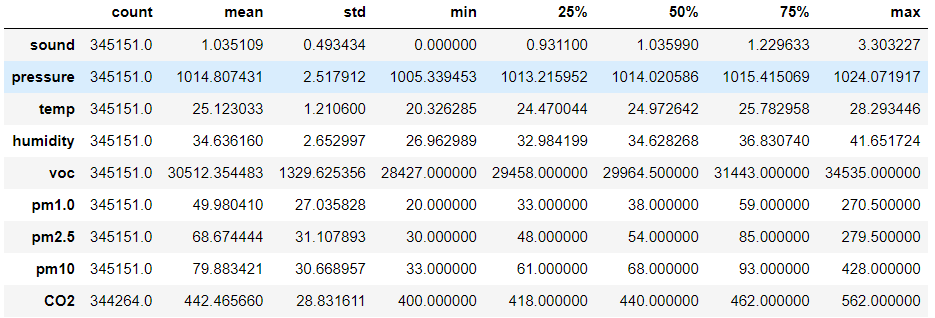
## Ток анализе података

У овом поглављу ће детаљније бити описана комплетна анализа над поменутим скупом података, укључујући и детекцију тачака промене (за појединачне атрибуте и за све атрибуте истовремено). Битно је напоменути да су у овом тексту издвојене и анализиране само најбитније ставке из практичног дела рада. Целокупан рад се налази у одговарајућој радној свесци.

### Визуелизација података

У овој секцији извршено је основно истраживање скупа података, као и визуелизација података.

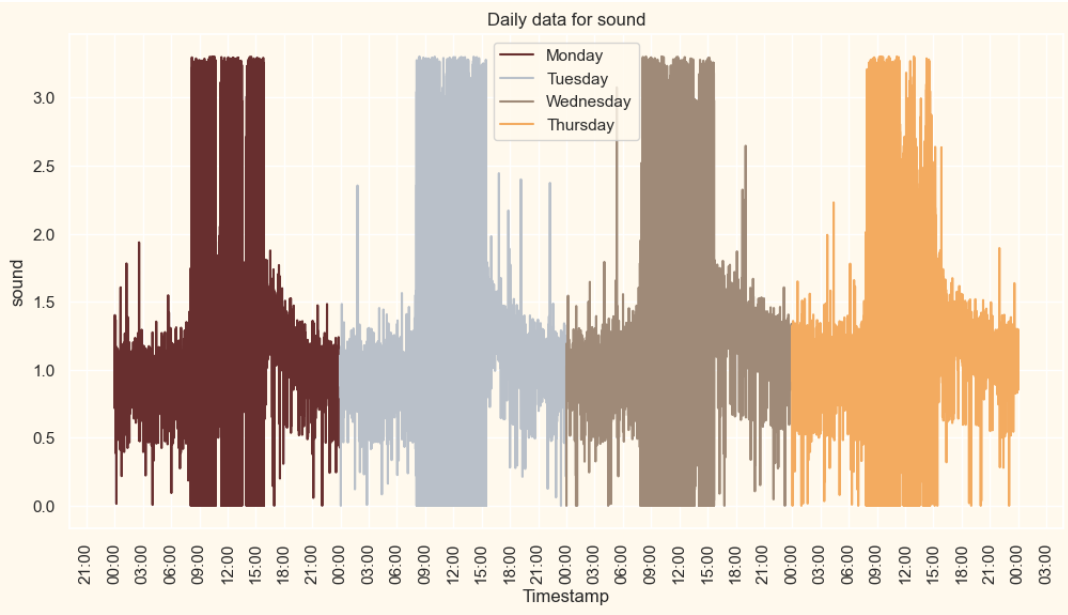
Најпре је извршено испитивање основних статистичких вредности унутар скупа. На следећој слици су те вредности приказане.



**Слика 2.3.1.1** Основне статистичке вредности

Из података можемо извести неке једноставне закључке о статистичким својствима различитих параметара. На пример, вредности звука се крећу у опсегу од 0 до 3.303 са стандардном девијациом од приближно 0.5. Такође, оно што је у овом моменту интересантно је вредност стандардне девијације за атрибут “*voc”* која је изузетно висока, што указује на велику варијансу међу подацима.

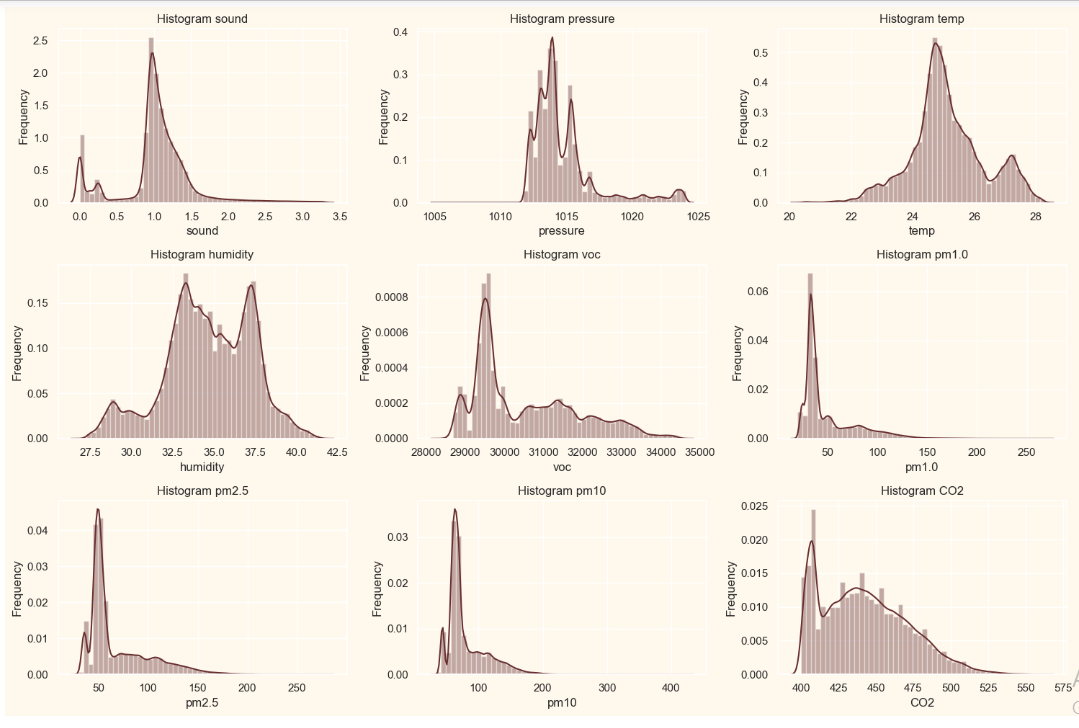
Након овога, врши се приказ дневних података за сва 4 радна дана. На слици су приказани дневни подаци само за један атрибут (*sound*), у радној свесци је могуће видети све приказе.



**Слика 2.3.1.2** Дневни подаци за звук

Као што се на слици може видети, звук достиже највеће вредности у периоду од 08:00 до 16:00. На основу овога, могуће је закључити да фабрика ради управо у том периоду и да је бука из тог разлога тада већа. Слично је урађено и за друге атрибуте, где се такође дошло до одговарајућих закључака о самим подацима. Више о свему овоме се налази у радној свесци практичног дела пројекта.

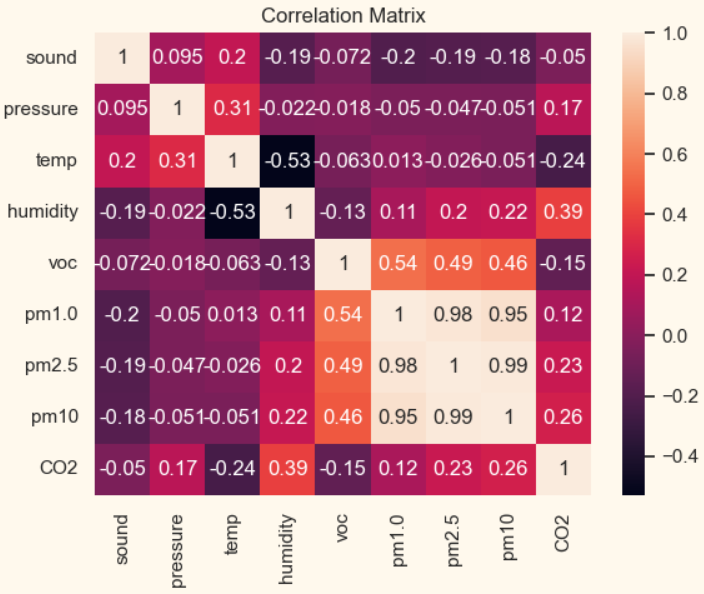
Затим, извршен је и приказ дистрибуције података.



**Слика 2.3.1.3** Дистрибуција података

Ниједан атрибут нема нормалну дистрибуцију података. Такође, примећујемо и постојање *outlier-*а.

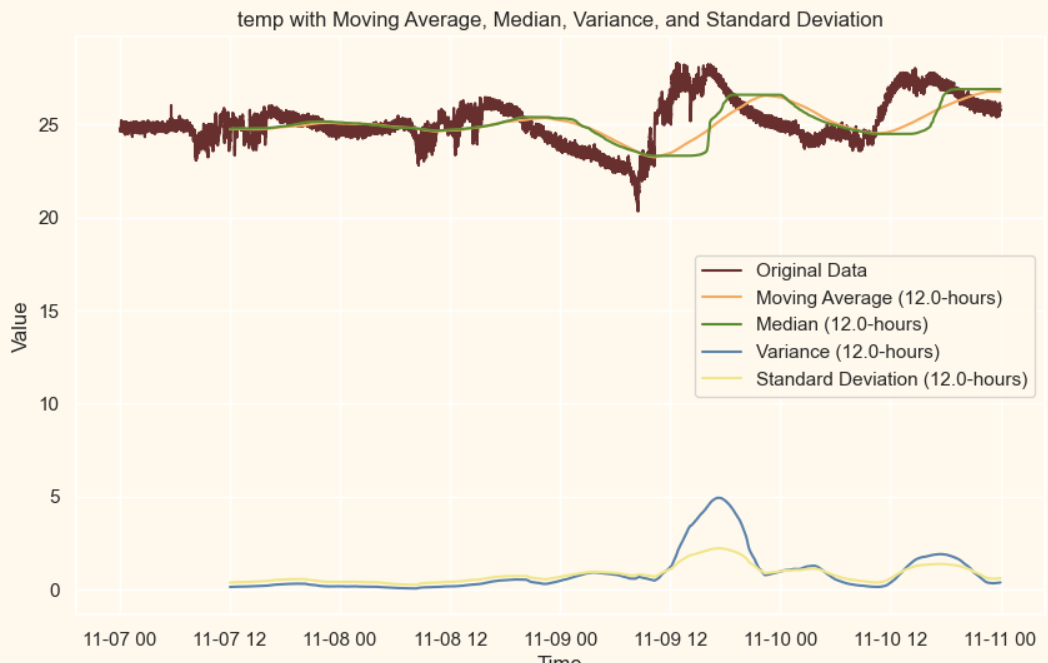
Након овога, приказана је матрица корелација. У овој матрици је уочено постојање веома високе корелације између сва три типа *pm* атрибута. Корелација је превазилазила 0.98 што се може и видети на слици.



**Слика 2.3.1.4** Корелациона матрица

Ова висока корелација је указала на то да је могуће извршити редукцију атрибута, што је касније у раду и одрађено стапањем ова три атрибута у један уз помоћ *PCA* алгоритма за редукцију димензионалности.

Даље, извршено је посматрање статистичких података у току времена (*over time*). Ово је такође визуелизовано за сваки од атрибута, али ће у овом теоријском делу бити приказан само за један (температура).



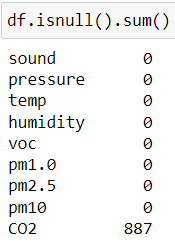
**Слика 2.3.1.5** *Moving average* за температуру

На слици можемо да видимо приказ позива имплементиране *moving average* функције. Као што се може приметити, вредности варијансе и стандардне девијације су изузетно ниске, док се код медијалних и средњих вредности не може уочити постојање тренда или сезоналности. На основу овог графика може се претпоставити да су подаци за температуру стационарни. Наравно, ово није довољно да би смо били сигурни, па ће касније бити извршена озбиљнија провера.

### Претпроцесирање података

У овој секцији примењене су неке основне методе чишћења и модификације самих података у циљу тачније анализе. Неке од метода које су примењене су: испитивање недостајућих података, дупликата, редукција атрибута и др.

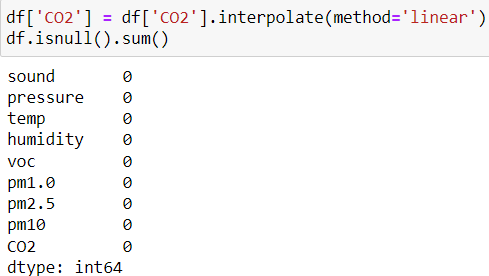
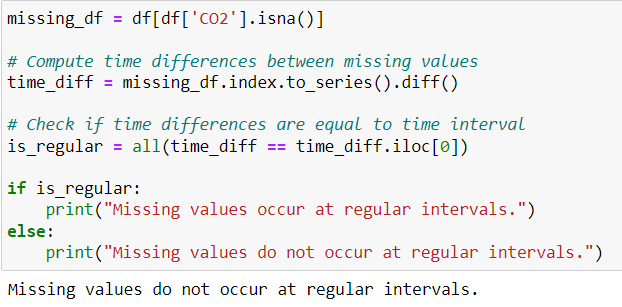
Најпре је извршена провера везана за недостајуће податке. Након ове провере, дошло се до закључка да постоје недостајући подаци и то конкретно за *CO2*, као што се и може видети на слици.



**Слика 2.3.2.1** Недостајући подаци

С обзиром да су у питању временски зависни подаци, није решење заменити ове податке неком једноставном методом попут замене вредности средњом вредношћу и сл.

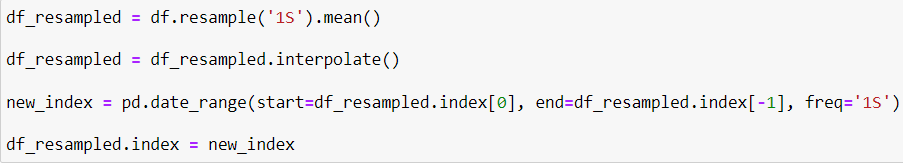
Затим је било потребно испитати да ли постоји одређени образац по којем подаци недостају. С обзиром да смо након испитивања установили да нема обрасца (не недостају у зависности од времена када су се десили), било је могуће употребити методу интерполације и на тај начин извршити допуну недостајућих вредности. Све ово можемо видети на следећим сликама.



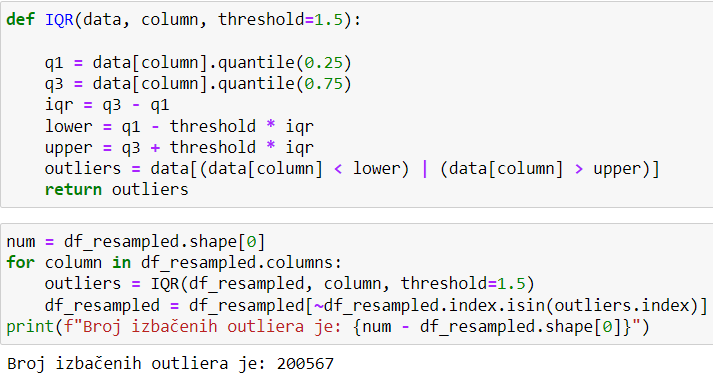
**Слика 2.3.2.2** Испитивање обрасца и интерполација

Након овога, извршена је провера да ли постоје дупликати у скупу података. Утврдило се да не постоје.

Следећи кораци били су промена фреквенције података (*resampling*) на фреквенцију од 1*Hz,* као и отклањање потенцијалних *outlier-*а. Ова два корака приказана су на следећим сликама.



**Слика 2.3.2.3** *Resampling*

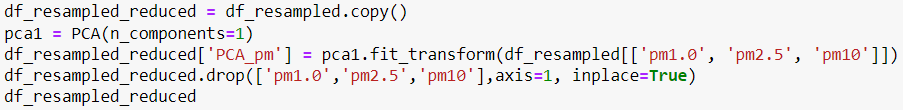


**Слика 2.3.2.4** Брисање *outlier-*а

Брисање *outlier-*а извршено је коришћењем најпопуларније методе за ову сврху, *IQR* методе. Као што можемо да видимо, избачено је преко 200 хиљада *outlier-*а што је заиста велики број.

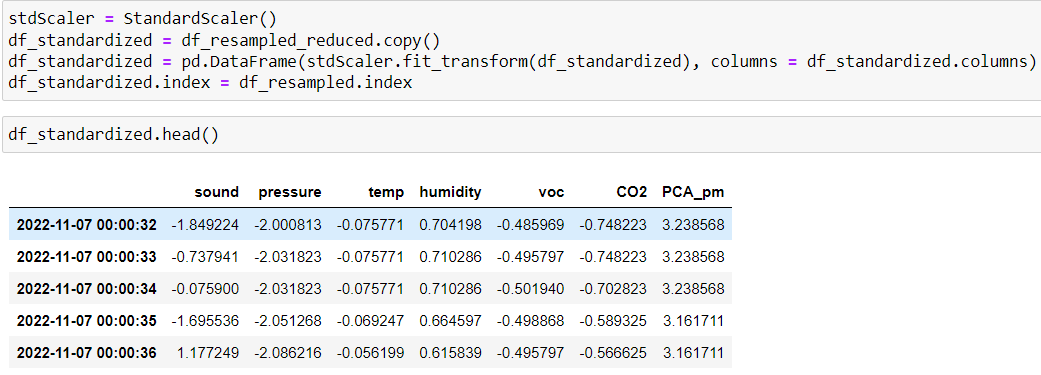
Затим, одрађене су две веома битне ставке. То су, малочас поменута, редукција атрибута и стандардизација скупа података.

Што се редукције атрибута тиче, коришћен је *PCA* алгоритам како би се три атрибута (*pm1.0, pm2.0, pm10*) стопили у један атрибут (*PCA\_pm*). Ево и приказа поменутог.



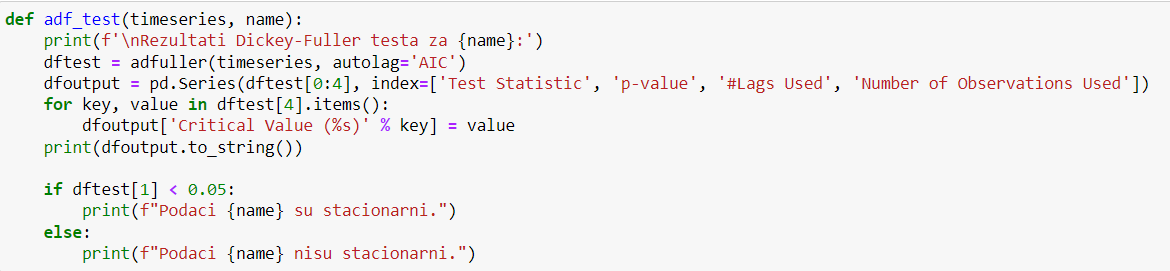
**Слика 2.3.2.5** *PCA*

За стандардизацију података коришћен је *StandardScaler.*



**Слика 2.3.2.6** Стандардизација података

На крају ове секције извршено је тестирање стационарности података коришћењем *Dickey-Fuller* теста. Након извршења овог теста, закључак је да су сви подаци стационарни, изузев података за притисак. Међутим, с обзиром на то да је *ruptures* библиотека отпорна на нестационарне податке, није било потребе пребацивати их у стационарне. Такође, у случају пребацивања у стационарне, дошло би до евентуалног губитка информација. На слици је приказана функција за тестирање.

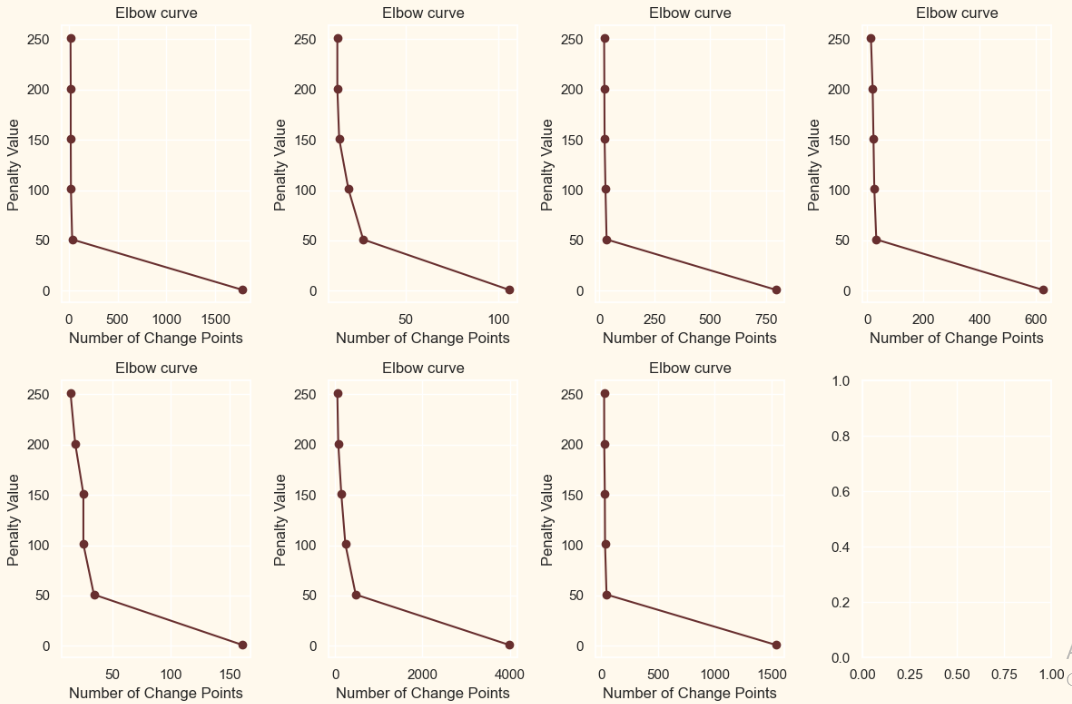


**Слика 2.3.2.7** *Dickey-Fuller* тест

### *Change point detection (Univariate)*

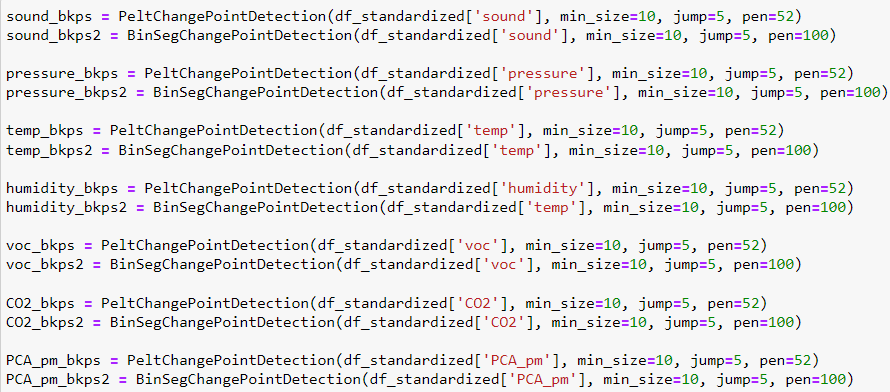
У овој секцији примењени су алгоритми који су нешто раније описани. Подсећања ради, у питању су алгоритми *PELT* и *BinSeg.* С обзиром да немамо унапред познат број тачака промене, неопходно је било користити управо ове алгоритме, али и на најлакши могући начин одредити оптималну вредност *penalty* променљиве, која је саставни део ових алгоритама.

У ову сврху, коришћена је ткзв. лакат крива. На *x* оси, налази се број детектованих тачака промене, док се на *y* оси налази вредност пенала. То у пракси изгледа овако.



**Слика 2.3.3.1** Лакат криве

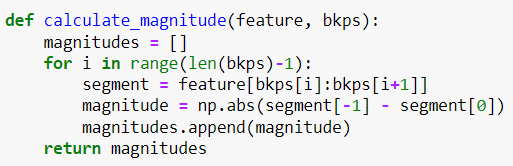
Као што се може приметити, оптимална вредност за пенал је између 50 и 60 за готово све атрибуте. Ја сам у овом раду користио вредност 52. Наравно, битно је напоменути да се ове лакат криве исцртавају само за случај *PELT* алгоритма. Разлог овога је брзина извршавања алгоритма. Наиме, *PELT* алгоритам има своју имплементацију у *C* програмском језику, док су остали алгоритми писани искључиво у *Python-*у. Због овога је брзина извршења овог алгоритма неупоредиво већа у односу на остале алгоритме. Због ограничених перформанси рачунара и брзине извршавања, определио сам се да лакат криву прикажем, и на основу ње изведем закључке, само за *PELT* алгоритам.



**Слика 2.3.3.2** *Univariate Change point detection*

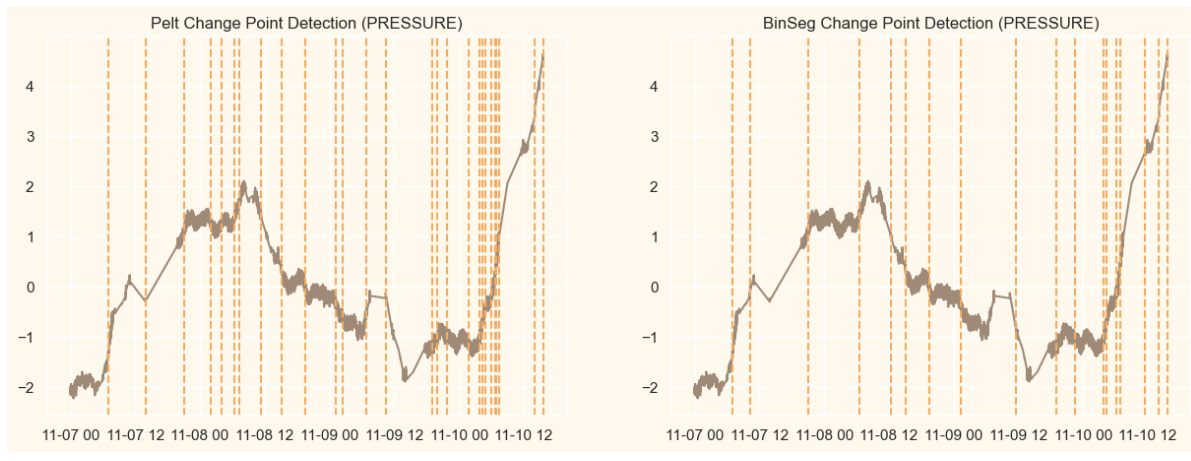
На слици 2.3.2.9 се налазе позиви одговарајућих алгоритама и рачунање тачака промене.

Након овога, извршено је рачунање магнитуда успона и падова у детектованим тачкама промене коришћењем следеће функције.

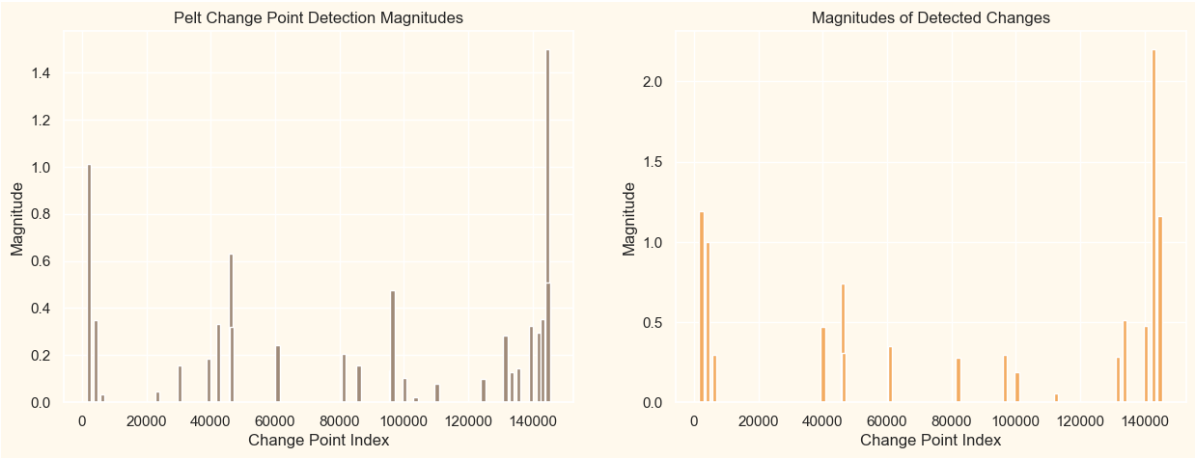


**Слика 2.3.3.3** Рачунање магнитуда тачака промене

Следи приказ тачака промене и њихових магнитуда за сваки од атрибута посебно.



**Слика 2.3.3.4** Приказ тачака промене



**Слика 2.3.3.5** Приказ магнитуда тачака промене

Магнитуде тачака промене нам дају информације о скоку, односно паду статистичких параметара у датом тренутку. С обзиром да не постоје *ground truth* вредности, ово може да послужи као добар индикатор оцене примењених алгоритама, што је касније и урађено.

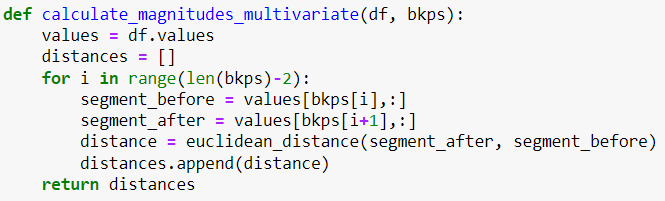
У конкретном случају, закључак је да се *PELT* алгоритам показао боље од *BinSeg* алгоритма, када узмемо у обзир перформансе и магнитуде.

### *Change point detection (Multivariate)*

У овом одељку је вршена *multivariate* детекција тачака промене над *enviroment* скупом података. С обзиром на обимност ове секције, у овом теоријском делу рада ће бити приказани само резултати најбоље показаног модела. Све пробане алгоритме и резултате могуће је видети у радној свесци практичног дела рада.

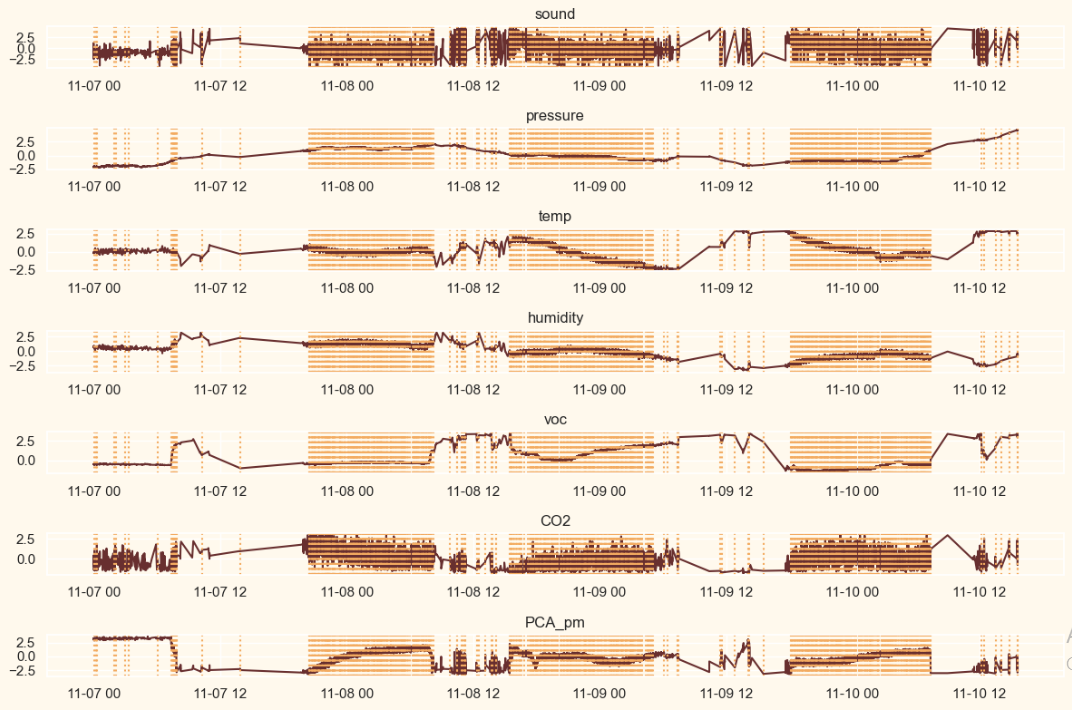
У овом делу, за разлику од малопре, рачунамо тачке промене када узмемо у обзир све атрибуте скупа података, тј. тачке промене су заједничке за све атрибуте. Управо ово је и био крајњи циљ овог рада.

За разлику од *univariate* анализе где су рачунате појединачне магинтуде, овде је неопходно упоређивати векторе у тачкама промене. У ову сврху коришћено је еуклидско растојање између тренутака у тачки промене и следећег тренутка. На овај начин постижемо сличан ефекат као малочас. Ово је приказано на следећој слици.

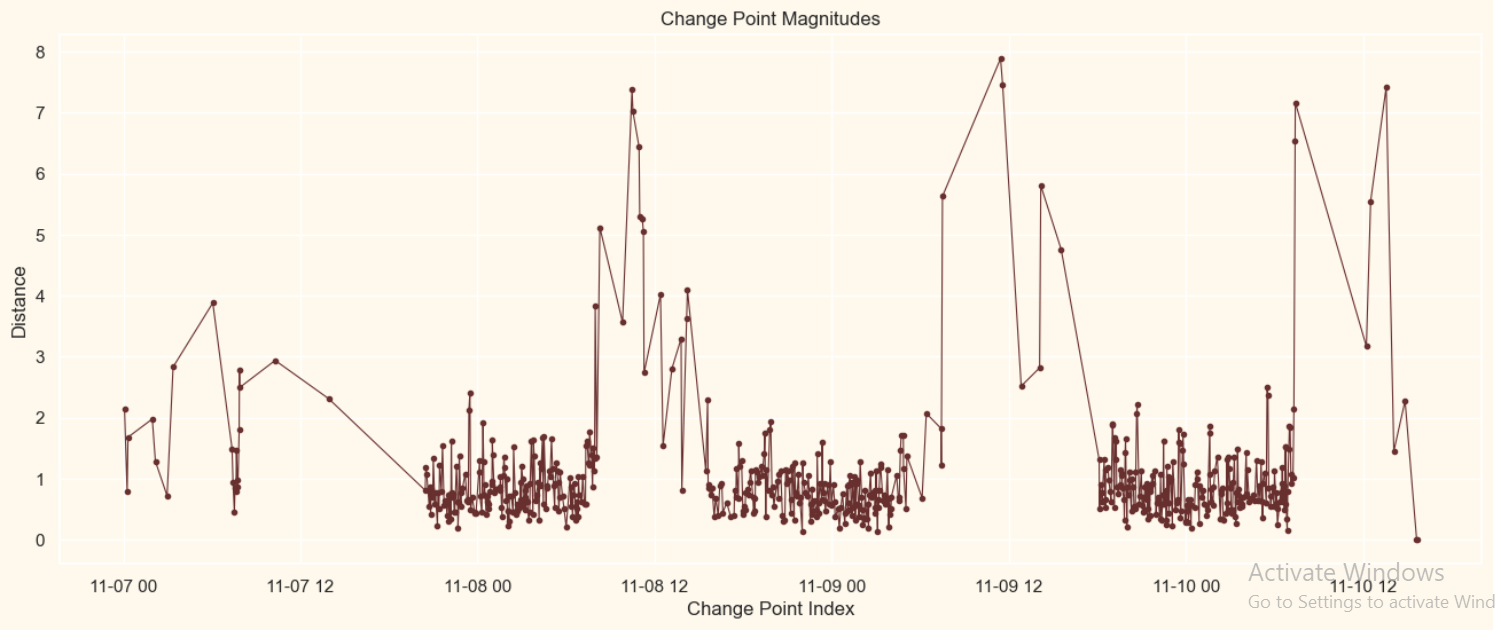


**Слика 2.3.4.1** Функција за рачунање еуклидског растојања у тачкама промене

Приказ добијених тачака промене и одговарајућих магнитуда за случај *PELT* алгоритма може се видети на слици испод.



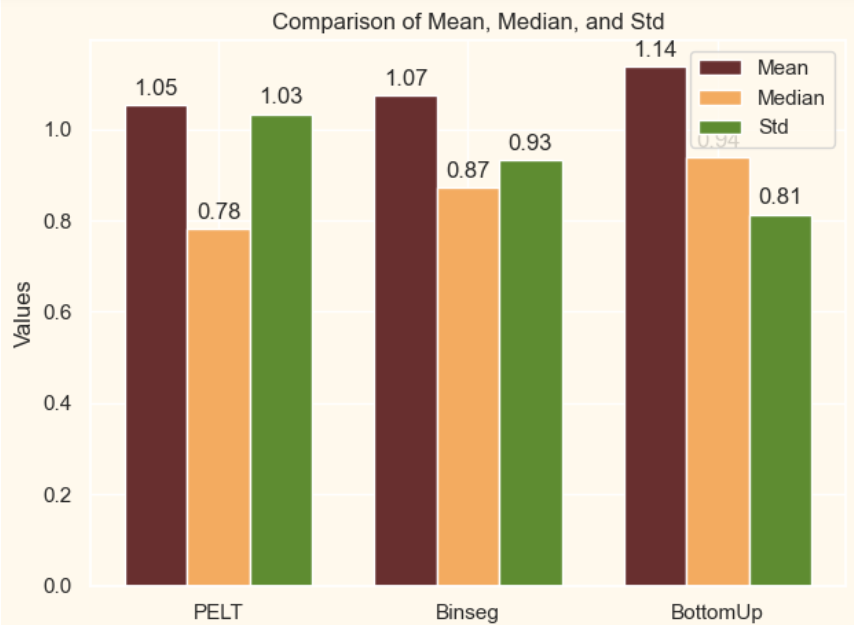
**Слика 2.3.4.2** Тачке промене (*Multivariate, PELT*)



**Слика 2.3.4.3** Магнитуде тачака промене (*Multivariate, PELT*)

Анализом графикона, могуће је закључити да је *PELT* алгоритам показао задовољавајуће перформансе у овој ситуацији. Овај закључак произлази из евалуације еуклидског растојања међу вредностима временских серија у тренутку детекције тачке промене и непосредно након ње. Већа вредност еуклидског растојања указује на значајнију промену у том тренутку. Имајући у виду учестале скокове и падове на графикону, може се закључити да се овај алгоритам показао адекватним за конкретну ситуацију.

На крају, имамо график који илуструје упоређивање свих одрађених алгоритама.



**Слика 2.3.4.4** Упоредна анализа

**Средња удаљеност**:

Алгоритам са мањом средњом удаљеношћу обично има тенденцију да у просеку детектује мање промене. У овом случају, *PELT* има нешто мању средњу удаљеност у односу на *BinSeg* и *BottomUp*. То указује на то да *PELT* можда најбоље детектује мање промене уопштено.

**Медијана удаљености**:

Медијана удаљености је мање осетљива на екстремне вредности у поређењу са средњом вредношћу. Нижа медијана удаљености сугерише да алгоритам доследно детектује промене мањег обима. Сва три алгоритма имају сличне медијане удаљености, што указује да сви релативно доследно детектују промене. Такође, најбоље се показао *PELT* алгоритам.

**Стандардна девијација (*std*)**:

Нижа стандардна девијација указује на то да су удаљености чвршће груписане око средње вредности, што сугерише већу доследност у детекцији. У овом случају, сва три алгоритма имају сличне стандардне девијације, што значи да је варијација у детектованим обимима промена упоредива.

Сходно свему наведеном, опредељујемо се за *PELT* алгоритам.

# Закључак

Овај рад пружа свеобухватну анализу скупа података који се тичу окружења и добијени су из сензора у фабрици. Почевши од основних закључака, па све до визуализације података кроз графичке приказе, истраживање је пажљиво изведено. Важна фаза истраживања обухватала је и претпроцесирање података, а овај корак је детаљно разрађен у теоријском оквиру рада. Претпроцесирање се односило на руковање недостајућим вредностима, смањење димензионалности, стандардизацију и сличне технике.

Након припремних корака, примењени су алгоритми *PELT*, *BinSeg* и *BottomUp* за анализу промена како у случају појединачних променљивих, тако и у случају више променљивих одједном. Детаљно су описани и приказани резултати добијени овим алгоритмима. На основу анализе резултата, закључак је да се *PELT* алгоритам истакао као најефикаснији и најпрецизнији у детекцији тачака промена, како у случају *univariate* детекције, тако и у случају *multivariate* детекције.

Крајње решење овог истраживања је и креирање новог *csv* фајла који садржи две колоне. Прва колона садржи временске ознаке (*timestamp*), док друга колона означава тачке промена (означено са 1 ако је тачка промена, односно 0 ако није). Овај корак има значајну практичну примену у даљим анализама и праћењу промена током времена.

У целини, овај рад представља дубоко истраживање и анализу података о окружењу, са пажљиво изведеним корацима обраде и применом алгоритама за детекцију тачака промена. Његови резултати пружају важан увид у динамику промена у окружењу и наглашавају значај *PELT* алгоритма за прецизну и поуздану анализу промена у подацима.

# Литература

1. *An Evaluation of Change Point Detection Algorithms, Gerrit J. J. van den Burg, Christopher K. I. Williams*
2. *A* *survey of methods for time series change point,* *Samaneh Aminikhanghahi, Diane J. Cook*
3. *https://github.com/deepcharles/ruptures/tree/master/docs*, (Приступљено 15.08.2023)