N-dimensional analysis pipeline

Comparison of different methods:

- Tasseography, shadow projections
- Multidimensional analysis

Multidimensional analysis using ND pipeline - current usage

Materialized view

Examples (Mostly backup slides)

- Space charge distortion calibration and studies
- PID calibration and monitoring
- Tracking performance characterization (TRD in combined tracking)
- TRD tracking optimization
- New event and track selection cuts (dEdx and chi2 QA analysis)
- Visualization development

Multidimensional analysis using ND pipeline - Future plans

- ROOT 6

Tasseography (also known as tasseomancy or tassology) is a divination or fortune-telling method that interprets patterns in tea leaves, coffee grounds, wine sediments or 1D histograms

Spring Pouchong tea (Chinese: 包種茶; pinyin: *Bāozhòngchá*) leaves that may be used for tasseography divination

An example of a tea leaf reading showing a dog and a bird on the side of the cup.

https://en.wikipedia.org/wiki/Tasseography

Example wrong statements:

"Detector noise did not change":

- **1D conclusion:** 1D mean and rms is "in range"

- **Reality could be:** relative increase of the noise in critical/noisy regions by factor 2-3 not spotted

"DCA resolution is fine":

- **1D conclusion:** TPC σ_{DCA} is 1 cm as usual

- **Reality could be:**

   - DCA resolution at high pt 3-4 times worse (3-4 mm instead of the 1 mm)
   - DCA is biased as function of phi

"TPC detector occupancy is outside of the range":

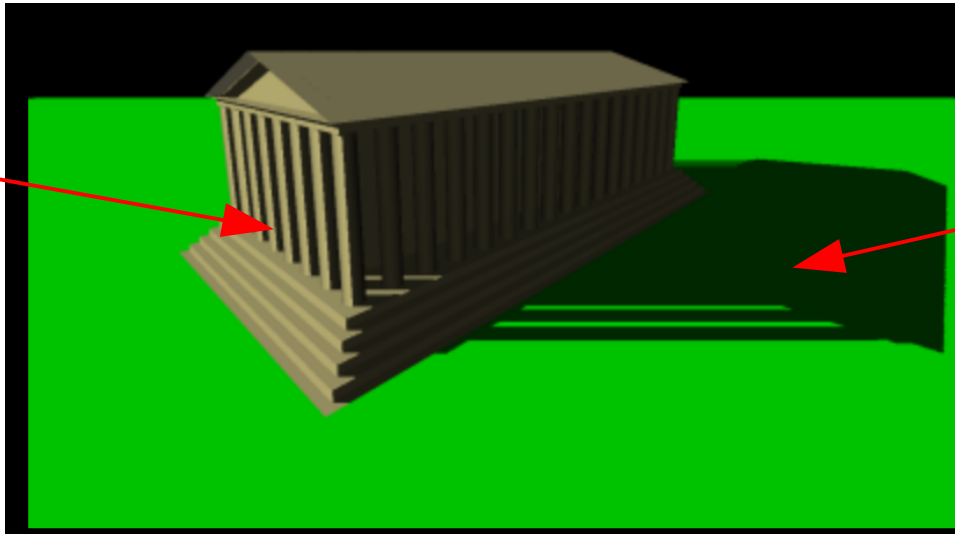- **1D conclusion:** Mean TPC occupancy> limit

- **Reality could be:** occupancy increased because of increase of IR resp. beam background

"MC describes the data":

- **1D conclusion:** Mean mass bias in MC has the same sign as data

- **Reality could be:** The sign of bias is pure coincidence (50 % probability). Pt dependence not reproduced at all …

Our object
E.g occupancy
map from
AMORE
(3D)

Our current
TPC DQM

$$\sigma_{\vec{A} \ominus \vec{A}_{ref}} \leq \sigma_{\vec{A}} (+) \sigma_{\vec{A}_{ref}}$$

Projections could be misleading
- https://youtu.be/a7LCTT7HKzc

Guessing from 2D projection more reliable than Tasseography

- some imagination to be involved (see next slides)
- Information lost

Alarms to be based on some invariance - e.g the difference between the object and referenced object

- after projection impossible
- in my typical cases variance σ_{A-Aref} is very often smaller by orders of magnitudes

$$\sigma_{\vec{A} \ominus \vec{A}_{ref}} \leq \sigma_{\vec{A}} (+) \sigma_{\vec{A}_{ref}}$$

Invariance/symmetries in N dimensions (A ref model vector):

- invariance in time (using e.g. reference run)
- invariance in space (e.g. rotation, mirror symmetry)
- data - physical model
- A side/C side, B field symmetry
- smoothness resp. local smoothness

Projections problems (hidden variables):

- **Information loss. Intrinsic spread of variable vectors A and A_{ref} is bigger than spread of A-A_{ref}**
  - gain maps, noise map, DCA bias, resolution maps, occupancy maps, sigma invariant mass maps ... as function of 1/pt, θ, occupancy, dEdx
- **Projected vector A depends on the actual distribution of hidden variable**
  - Sometimes misleading results
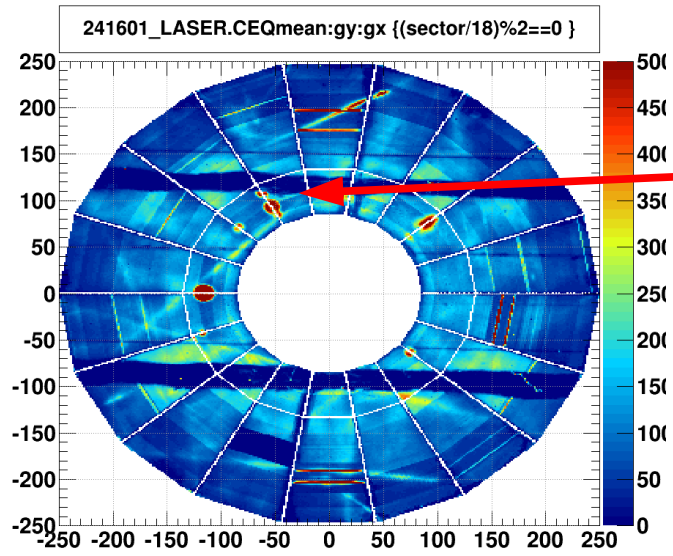 - Non trivial interpretation of projected observation

$$\sigma_{\vec{A}-\vec{A}_{ref}} < \sigma_{\vec{A}}(+)\sigma_{\vec{A}_{ref}}$$

Ion deposited at **Central electrode (CE)** decrease work function (online calibration/AMORE data as input)
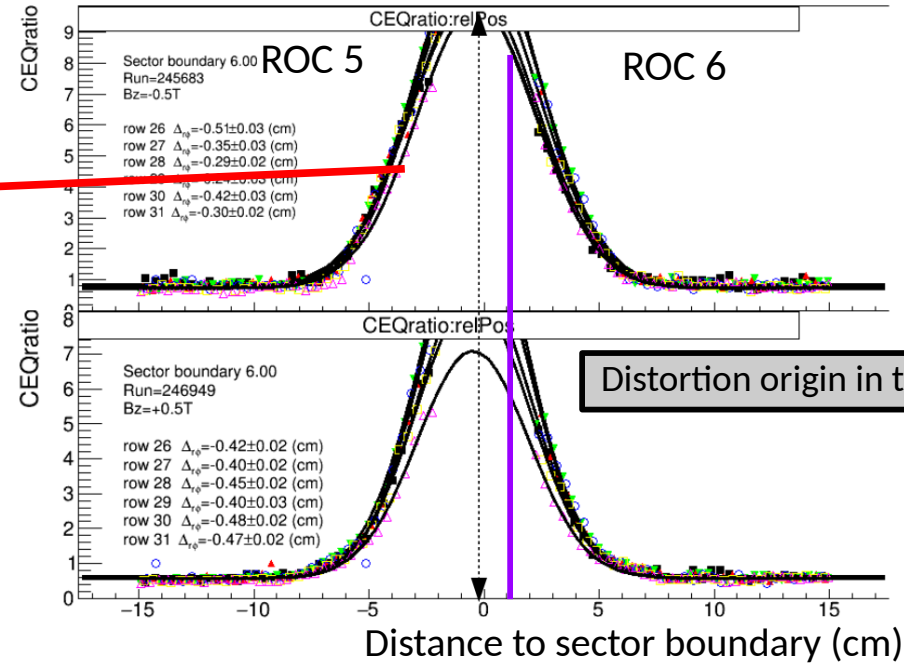
- → Increased emission of electrons during laser shots
- Message: Center of gravity fit of higher charge (distortion) at the sector gap

CE mean charge map normalized to reference data to cancel intrinsic dispersion

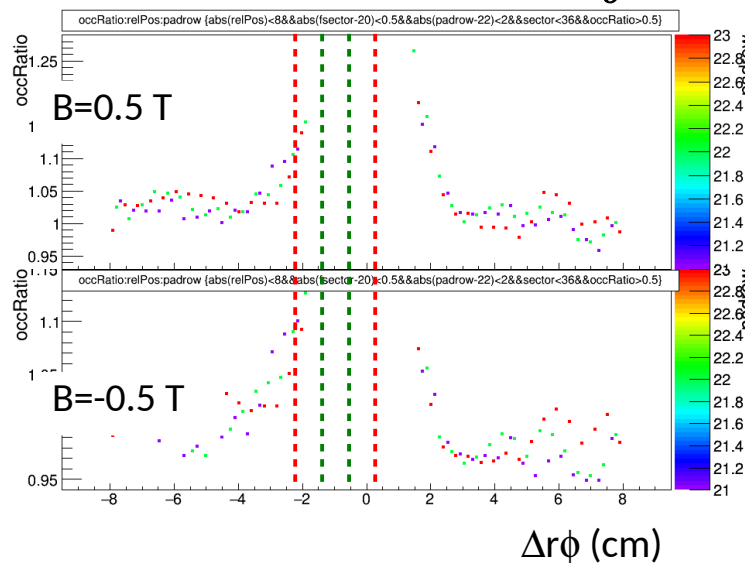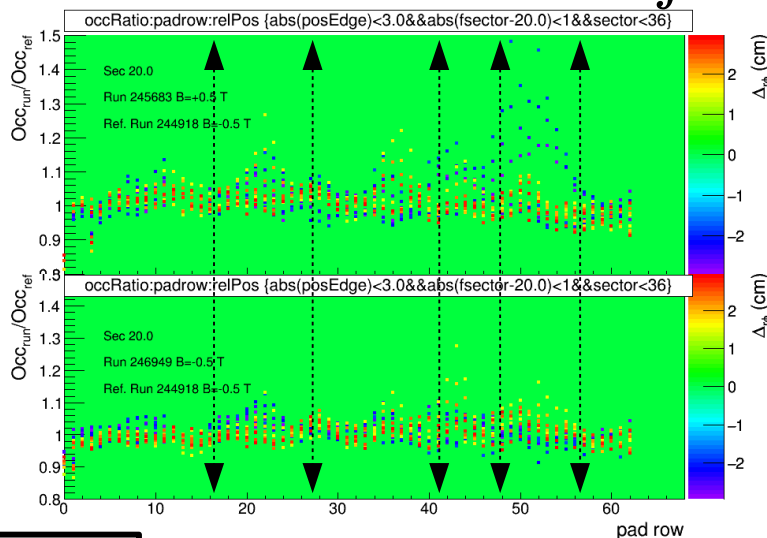$$R(\Delta r\phi, row) = \frac{Q_{CE}(\Delta r\phi, row)}{Q_{CEref}(\Delta r\phi, row)}$$

CE signal ratio - data/reference data

241601_LASER.CEQmean:gy:gx {(sector/18)%2==0 }

ROC 5   ROC 6

CEQratio:relPos

Sector boundary 6.00
Run=245683
Bz=-0.5T

row 26 $\Delta_{r\phi}$=-0.51±0.03 (cm)
row 27 $\Delta_{r\phi}$=-0.35±0.03 (cm)
row 28 $\Delta_{r\phi}$=-0.29±0.02 (cm)
row 30 $\Delta_{r\phi}$=-0.42±0.03 (cm)
row 31 $\Delta_{r\phi}$=-0.30±0.02 (cm)

CEQratio:relPos

Sector boundary 6.00
Run=246949
Bz=+0.5T

row 26 $\Delta_{r\phi}$=-0.42±0.02 (cm)
row 27 $\Delta_{r\phi}$=-0.40±0.02 (cm)
row 28 $\Delta_{r\phi}$=-0.45±0.02 (cm)
row 29 $\Delta_{r\phi}$=-0.40±0.03 (cm)
row 30 $\Delta_{r\phi}$=-0.48±0.02 (cm)
row 31 $\Delta_{r\phi}$=-0.47±0.02 (cm)

Distortion origin in the gap

Distance to sector boundary (cm)

$$\sigma_{\vec{A} \ominus \vec{A}_{ref}} \leq \sigma_{\vec{A}} (+) \sigma_{\vec{A}_{ref}}$$

Model:

$$\frac{N_{Cl}(IR)}{N_{Cl}(IR=0)} = \overline{\frac{(w+(\Delta_{r\phi}(r_\phi+w/2)-\Delta_{r\phi}(r_\phi-w/2)))}{w}}$$

$$R = \left(\frac{Occ}{<Occ_{ROC}>}\right)_{IR} / \left(\frac{Occ}{<Occ_{ROC}>}\right)_{IR=0}$$

$$\overline{Z} \approx 125cm$$

Conclusion: Distortion origin in the gap

Increase of occupancy close to the hot-spot region due to space-charge distortion
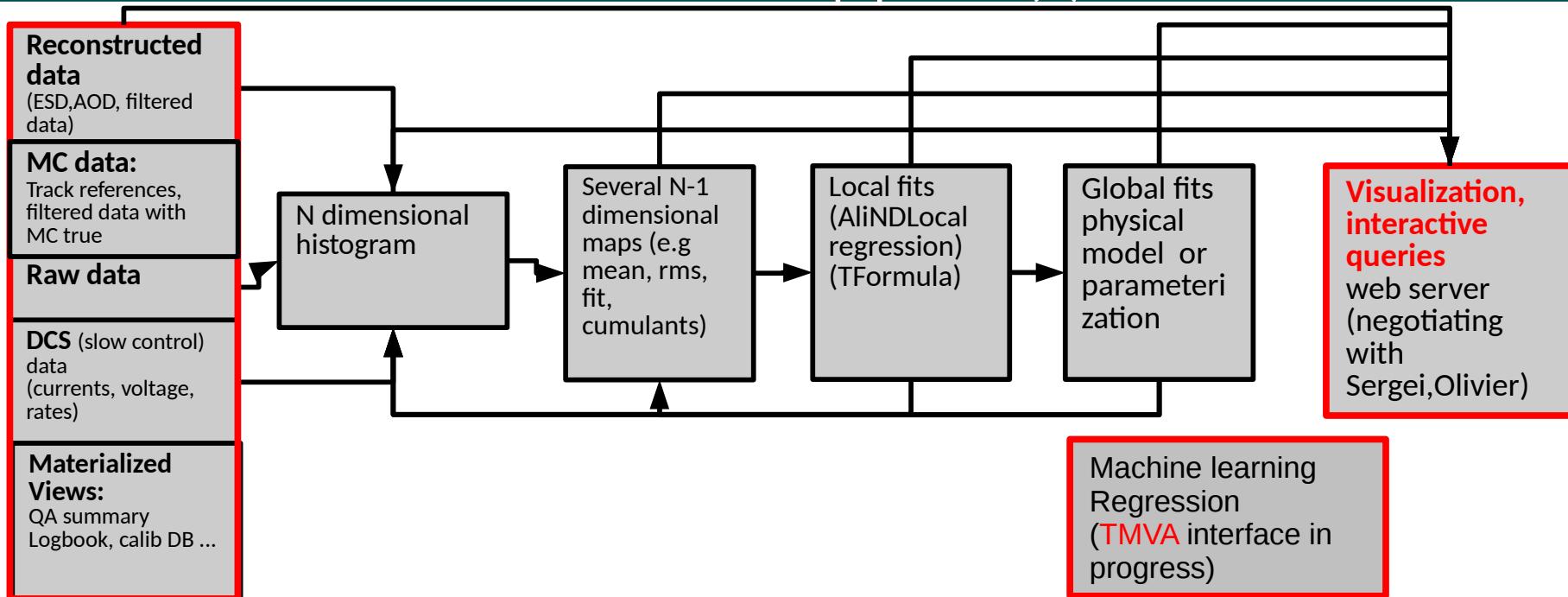Precise measurement of distortion origin - **derivative of distortion** with sub-pad granularity →
Distortion origin well localized
Without normalization to reference effect is invisible (first analysis)

Multidimensional analysis pipeline

Reconstructed data
(ESD,AOD, filtered data)

MC data:
Track references, filtered data with MC true

Raw data

DCS (slow control) data (currents, voltage, rates)

Materialized Views:
QA summary Logbook, calib DB …

N dimensional histogram

Several N-1 dimensional maps (e.g mean, rms, fit, cumulants)

Local fits (AliNDLocal regression) (TFormula)

Global fits physical model or parameterization

Visualization, interactive queries web server (negotiating with Sergei,Olivier)

Machine learning Regression (TMVA interface in progress)

$$f(p_0, p_1, p_2, \ldots) \neq f_0(p_0) \oplus f_1(p_1) \oplus f_2(p_2) \oplus \ldots$$

Standard calibration/performance maps and QA done and interpreted in multidimensional space

- dimensionality depends on the problem to study (and on available resources)
- Data →Histogram → set of ND maps → set of NDlocal fits → Global fits
  - **for all objects standard generic support of TFormula and TTreeFormula**
  - combining different maps using interpreted (or compiled) functions

Why N-dimensions ?

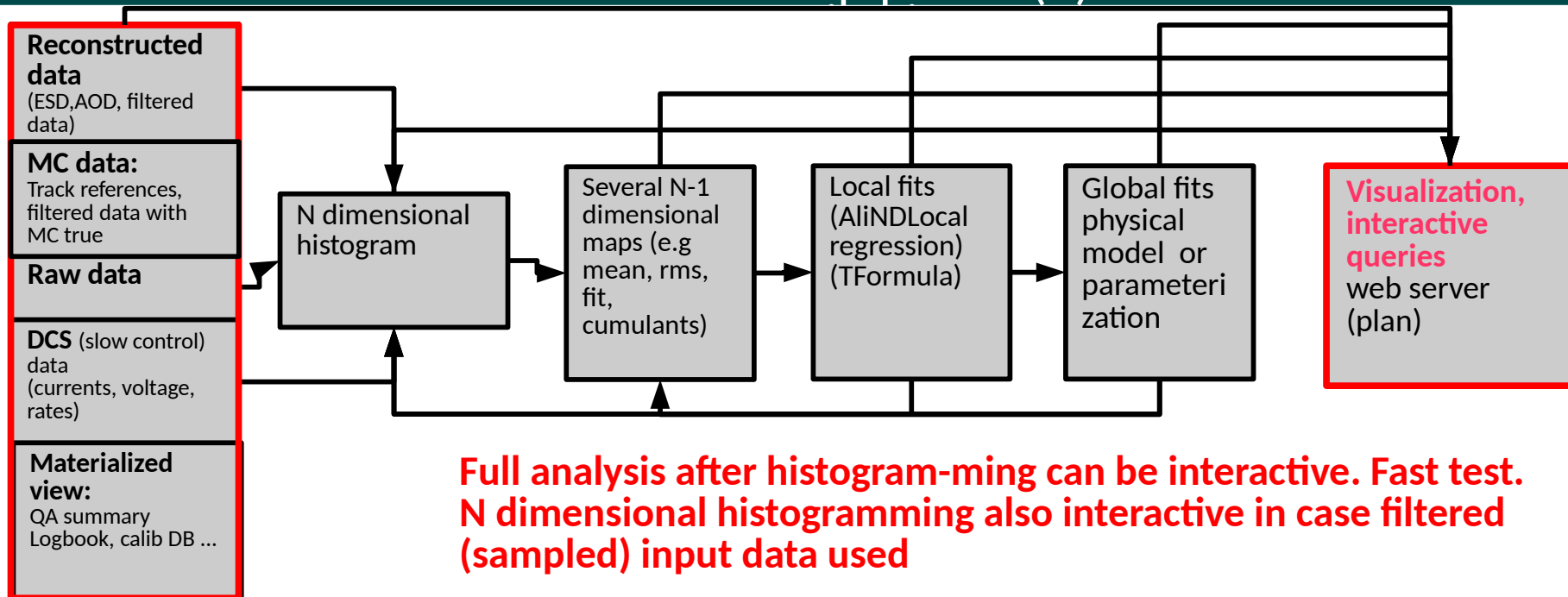- **Physics is in n-dimensions - dimensions do not factorize**

Reconstructed data
(ESD, AOD, filtered data)

MC data:
Track references, filtered data with MC true

Raw data

DCS (slow control) data (currents, voltage, rates)

Materialized views:
QA summary Logbook, calib DB …

N dimensional histogram

Several N-1 dimensional maps (e.g mean, rms, fit, cumulants)

Local fits (AliNDLocal regression) (TFormula) … ML regression

Global fits physical model or parameterization

Visualization, interactive queries web server (in progress)

Generic code. Minimizing amount of custom macros.
Standardizing ND analysis/functional representation
"Declarative" programming - simple queries

libSTAT - Pipeline of standalone tools (individual tools presented before within ALICE OFFLINE weeks)

- N dimensional histogramming
  - THn/THnSparse - part of root
  - AliTreePlayer::MakeHistogram ….
- THn → Map (tree)
  - part of the TStatToolkit
- Map(Tree) → Local fits
  - AliNDLocalRegression, TMVA+TMVA::Reader interface
- Map(Tree) → Global fits (physical models, parameterizations)
  - AliTMinuitToolkit

Reconstructed data (ESD,AOD, filtered data)				
MC data: Track references, filtered data with MC true	N dimensional histogram	Several N-1 dimensional maps (e.g mean, rms, fit, cumulants)	Local fits (AliNDLocal regression) (TFormula)	Global fits physical model or parameteri zation
Raw data				
DCS (slow control) data (currents, voltage, rates)				Visualization, interactive queries web server (plan)
Materialized view: QA summary Logbook, calib DB …				

Full analysis after histogram-ming can be interactive. Fast test. N dimensional histogramming also interactive in case filtered (sampled) input data used

Visualization/query - generic support exist (libSTAT)

- TFormula (TF1, TTreeFormula, AliTreeFormula),TStatToolkit (Metadata)
- AliTreePlayer
- AliPainter and AliDrawStyle(CSS)

New extensions:

- CSS library

Declarative (not yet parallel) analysis Integration with TDataFrame (ongoing)
- e.g THn interface in new Release

- Multidimension drawing support (Histogram, TFormula, TTree)
  - interface inspired by some python libraries, reporting
- Web server (static and jsroot based)

Macro

```
void LoadChain(const char *filteredList="filtered.list",Int_t nChunks=1);
void MakeHistograms(Int_t nEvents=100000);
void MakeMaps();
void makeNDLocalFit(TString varName, TString customCut, TString errorVar);
```

Define variables and derived variable (e.g using ND parameterization)

```
treeA->SetAlias("ratioTotMax0", "fTPCdEdxInfo.GetSignalTot(0)/fTPCdEdxInfo.GetSignalMax(0)");
treeA->SetAlias("ratioTotMax1", "fTPCdEdxInfo.GetSignalTot(1)/fTPCdEdxInfo.GetSignalMax(1)");
treeA->SetAlias("pullTotMax0","(ratioTotMax0-hisRatioTotMax0Dist.meanGFit)/hisRatioTotMax0Dist.rmsGFit");
treeA->SetAlias("pullTotMax1","(ratioTotMax1-hisRatioTotMax1Dist.meanGFit)/hisRatioTotMax1Dist.rmsGFit");
```

MakeHistograms

```
s+= "ratioTotMax0:atgl:multA:pileUpZ:sdEdxMax:#isSelected>>hisRatioTotMax0(100,0.5,2.0,20,0,1,20,0,20000,5,-
50,50,20,6,32);";
s+= "ratioTotMax1:atgl:multA:pileUpZ:sdEdxMax:#isSelected>>hisRatioTotMax1(100,0.5,2.0,20,0,1,20,0,20000,5,-
50,50,20,6,32);";
AliTreePlayer::MakeHistograms(treeQ, s, "nclCut&&chi2Cut&&downscale&&abs(qPt)<1.25&&nclCutGold&&pileUpCut", 0,
nEvents, -1, 15);
```

makeNDLocalFit (local regression + register fit functor for later usage)

```
makeNDLocalFit("hisRatioTotMax1Dist.meanG","","sqrt(0.01+hisRatioTotMax1Dist.rmsG**2/hisRatioTotMax1Dist.entries)");
makeNDLocalFit("hisRatioTotMax2Dist.meanG","","sqrt(0.01+hisRatioTotMax2Dist.rmsG**2/hisRatioTotMax2Dist.entries)");
```

makeGlobalFit (dEdx resolution example)

```
AliTMinuitToolkit *fitterResolMaxTot = new AliTMinuitToolkit("AliTMinuitToolkitFitterResolMaxTot.root");
formulaResolMaxTot = new TFormula("formulaResolMaxTot", "sqrt( x[0]**2 + ([0])*(x[1]^([1]))*(x[2]^([2])) )");
fitterResolMaxTot->FillFitter(treeMap,"rmsMaxTot0" ,"rmsTot0NormFit0:(multAMean/10000.):(1/
((sdEdxMaxCenter**2)*sqrt(1+atglCenter**2)))" , "fitCut1&&outlierCut", 0, 10000000);
fitterResolMaxTot->Bootstrap(20, "report1");
treeMap->SetAlias("rmsMaxTot0.Fit", fitterResolMaxTot->GetFitFunctionAsAlias().Data());
```

```
makeNDLocalFit("hisRatioMax01Dist.rmsG","isOK","sqrt(0.01**2+hisRatioMax01Dist.rmsG**2/hisRatioMax01Dist.entries)");
makeNDLocalFit("hisRatioMax12Dist.rmsG","isOK","sqrt(0.01**2+hisRatioMax12Dist.rmsG**2/hisRatioMax12Dist.entries)");
makeNDLocalFit("hisRatioMax02Dist.rmsG","isOK","sqrt(0.01**2+hisRatioMax02Dist.rmsG**2/hisRatioMax02Dist.entries)");
treeMap->SetAlias("rmsMax0","sqrt((hisRatioMax01Dist.rmsGFit**2+hisRatioMax02Dist.rmsGFit**2-
hisRatioMax12Dist.rmsGFit**2)/2.)");
TFormula *formulaResol0 = new TFormula("formulaResol", "[0]*(x[0]^abs([1]))*(x[1]^abs([2]))");
fitterResol0->SetFitFunction(formulaResol0, kTRUE;
fitterResol0->FillFitter(treeMap, "rmsMax0", "(1./sdEdxMaxMean**2):(1./sqrt(1+atglMean**2))",
                          "fitCut0&&outlierCut", 0, 10000000);
fitterResol0->Bootstrap(20, "report0");
treeMap->SetAlias("rmsMax0Fit0", fitterResol0->GetFitFunctionAsAlias().Data());
```

$$RMS_{Qi} = \sqrt{RMS^2_{Qi/Qj} + RMS^2_{Qi/Qk} - RMS^2_{Qi/Qk}/2}$$

$$RMS_{ROC} \times \sqrt{N_{CR}} \approx p_0 \left(dEdx^{p_1} \times \sqrt{(1+\tan(\theta))^2}^{p_2} \right)$$

$$(1)$$

Derived information: Relative intrinsic resolution estimated from the RMS of the dEdx ratios

Model: At low multiplicity relative dEdx resolution - power law scaling with the dEdx and tracklet length

- Parameters p_0, p_1 and p_2 fitted in global fits
  - fit using AliTMinuitToolkit + bootstrap method
  - error estimator as obtained in bootstrap

rmsMax0NormFit0:sdEdxMaxCenter:atglCenter {fitCut0&&multABin==1}

rmsMax0NormLF:rmsMax0NormFit0:atglCenter {fitCut0&&multABin==1}

$$RMS_{Qi} = \sqrt{RMS^2_{Qi/Qj} + RMS^2_{Qi/Qk} - RMS^2_{Qi/Qk}/2}$$

$$(1)$$
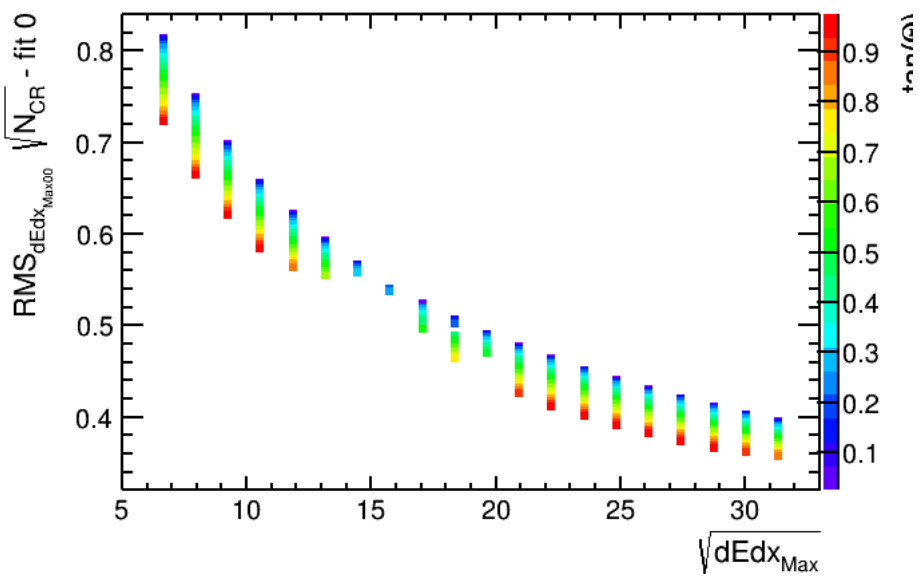
$$RMS_{ROC} \times \sqrt{N_{CR}} \approx p_0 \left(dEdx^{p_1} \times \sqrt{(1 + \tan(\theta))^2}^{p_2} \right)$$

Relative Intrinsic resolution for IROC as parameterized by global resp. local fit

At low IR agreement between dEdx resolution and power low model

- p0 =1.96 +-0.01
- p1=0.233+-0.001
- p2=0.40+-0.02

New functionality to precalculate/cache derived information:

- **AliESDtools** e.g:
  - Mean derived event properties
  - TRD extrapolation properties (chamber, position)
  - ITS/TPC standalone track matches
  - Recalculate dEdx - using up to data calibration
    - e.g wighting using resolution from previous slide
- **AliMCtools** e.g.:
  - TRD extrapolation properties (chamber, position)
  - Get closest reference
  - Get closest point

Reconstruction experts could provide additional derived information for the following queries

Materialized view and AliExternalInfo.
Example usage.

a materialized view (snapshot) is a database object that contains the results of a query, e.g

- a local copy of data located remotely
- a subset of the rows and/or columns of a table or join result
- a summary using an aggregate function

How?

- In DBMS the query result is cached as a concrete ("materialized") table that may be updated from the original base tables **from time to time**.

Why?

- Performance reasons, optimization:
- much **more efficient access**, at the **cost of extra storage** and of **some data being potentially out-of-date**.

When?

- Whenever frequent queries of the underlying databases are expensive
- Data warehouse scenarios

Goal

- Provide data for **Online analytical processing (OLAP)**, allowing for complex analytical and ad hoc queries with a rapid execution time

TTree interface to different sources of information (sources configured in configuration macro)

- Default configuration file $ALICE_ROOT/STAT/Macro/AliExteranlInfo.cfg

- Data cached locally: $AliExternalInfoCache

Default data sources:

Detector QA:

- QA.<Det>location e.g http://aliqatpc.web.cern.ch/aliqatpc
- QA.rawTPC.location http://aliqatpc.web.cern.ch/aliqatpc

Logbook:

- Logbook.location http://aliqamod.web.cern.ch
- Logbook.detector.location http://aliqamod.web.cern.ch
- TriggerAlias.location http://aliqamod.web.cern.ch
- TriggerClasses.location http://aliqamod.web.cern.ch

- Monalisa:

- MonALISA.ProductionCycle.location https://alimonitor.cern.ch/prod/?t=1&res_path=mif
- MonALISA.ProductionCPass.location https://alimonitor.cern.ch/prod/?t=3&res_path=mif
- MonALISA.ProductionMC.location https://alimonitor.cern.ch/prod/?t=2&res_path=mif
- MonALISA.ProductionCycleID.location https://alimonitor.cern.ch/prod/jobs.jsp?t=%s&res_path=mif
- MonALISA.MC.location https://alimonitor.cern.ch/MC/?res_path=mif
- MonALISA.RCT.location https://alimonitor.cern.ch/configuration/index.jsp?partition=%s&pass=%s&res_path=mif

```
Query:
AliExternalInfo info;
treeCalib = info.GetChain("QA.rawTPC","LHC18*","cpass1_pass1","QA.TPC;QA.EVS;Logbook");
treeCalib->SetAlias("separationPower","2*(meanMIPele-meanMIP)/
(resolutionMIP*meanMIP+resolutionMIPele*meanMIPele)");
treeCalib->SetAlias("qaPIDOK","resolutionMIP>0&&resolutionMIPele>0");
treeCalib-
>Draw("separationPower:interactionRate:qmaxQA.Sum()/72.","run>284700&&qmaxQA.Sum()/72.>30&&qaPIDOK","colz")
```


separationPower:interactionRate:qmaxQA.Sum()/72. (run>284700&&qmaxQA.Sum()/72.>40&&qaPIDOK)

2018
<Qmax> 48-57
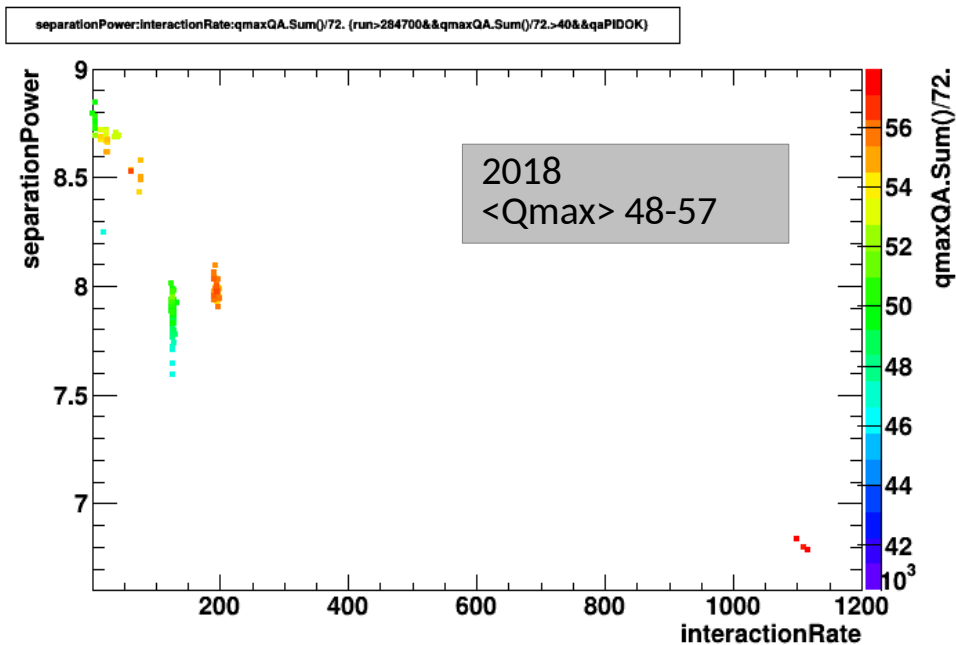
Simple example

Separation power (source):

- the gain (QA.rawTPC)
- interaction rate (QA.EVS)
- qmaxQA (Amore)
- B field (Logbook)

Alternative ? Slow and error prompt
Loop over runs.
    Open OCDB
        fit mean charge
    Open QAresouls.root
        fit electron, and MIP peak

2017 PID data analysis
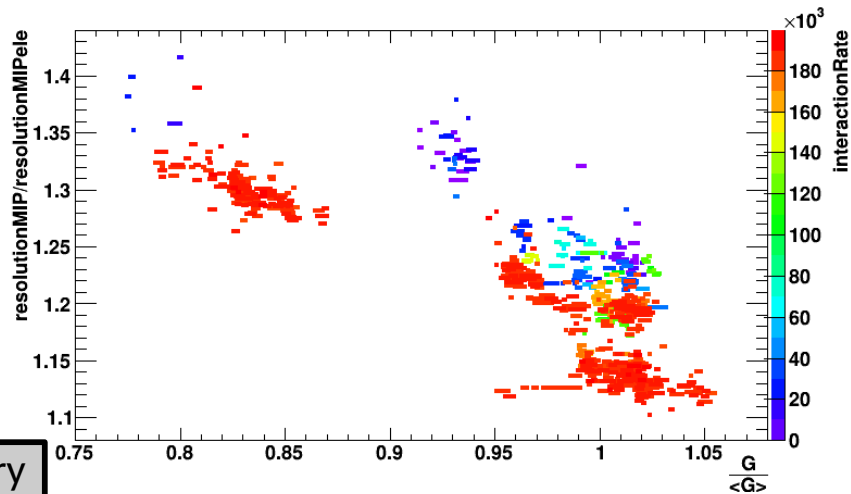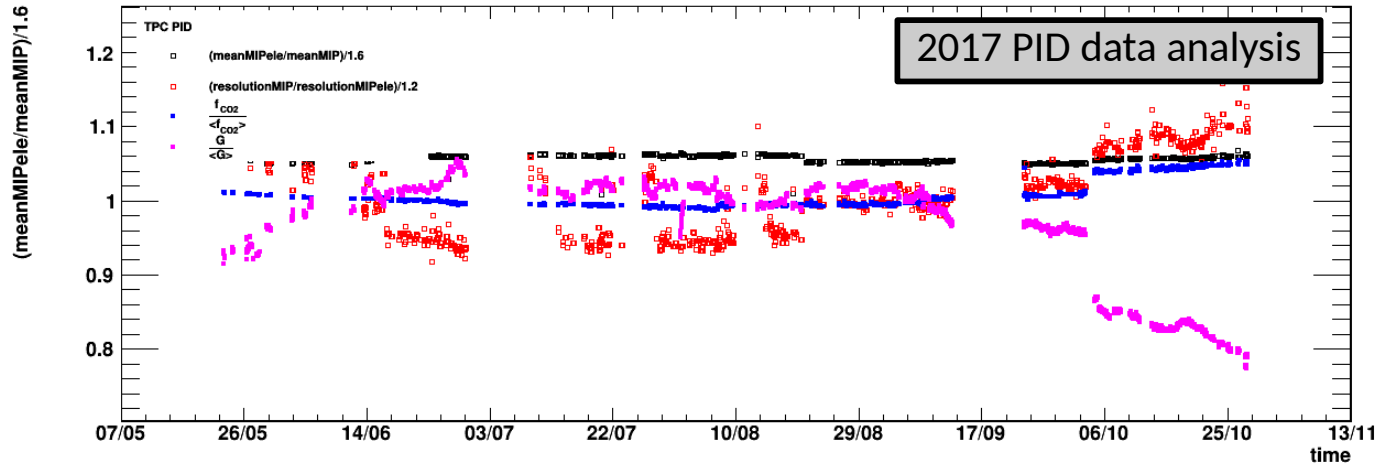
dEdx performance depends on several parameters (gain, IR, attachment..)
- different structures
- MIP resolution more affected by IR

~4%   relative increase of CO2
~20% modification of gain
~20% of worsening of the MIP resolution
~ 2 % modification of the El/MIP position
    BB modification

Query

```
treeCalib=externalInfo->GetChain("QA.rawTPC", period, pass,
"QA.TPC;QA.TRD;QA.ITS;QA.EVS;Logbook;Logbook.detector:TPC:detector==\"TPC\"");
graph = TStatToolkit::MakeMultGraph(treeCalib, "",  "(meanMIPele/meanMIP)/1.6;
(resolutionMIP/resolutionMIPele)/1.2;co2_Norm;CGainMIP_Norm:time","runCut&&CGainMIP>
0&&interactionRate<300000", "25;25;21;21","1;2;4;6", 0,0.6,3,legend);
```
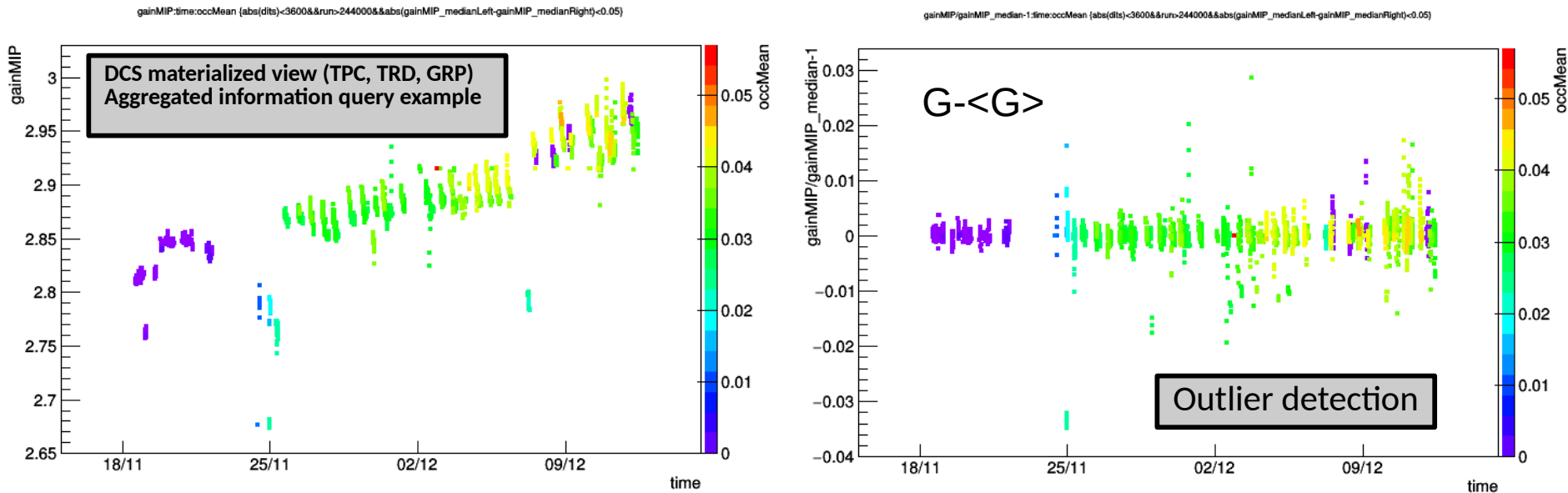
Left: Gain calibration as function of time (O(min)) time intervals

- Interaction rate modulation due imperfect ion tail/common mode correction (without Iontail/common mode correction O(20-30%) decrease of Q per central event

Right:

- Gain - local median gain (+-20 min). Outlier time intervals O(+-2%) visible
- Could be problematic for correlation studies

```
AliTreePlayer::AddStatInfo(treeCalib, treeCalib, "time:gainMIP", 5*10*60,
"median:medianLeft:medianRight:rms:mean");
treeCalib->Draw("gainMIP/gainMIP_median-
1:time:gainMIP","abs(dits)<3600&&run>244000&&abs(gainMIP_medianLeft-
gainMIP_medianRight)<0.05","colz");
```

`ntrITSRatio:T:sectorBin {entries>50&&abs(sectorBin-2)<1.1}`
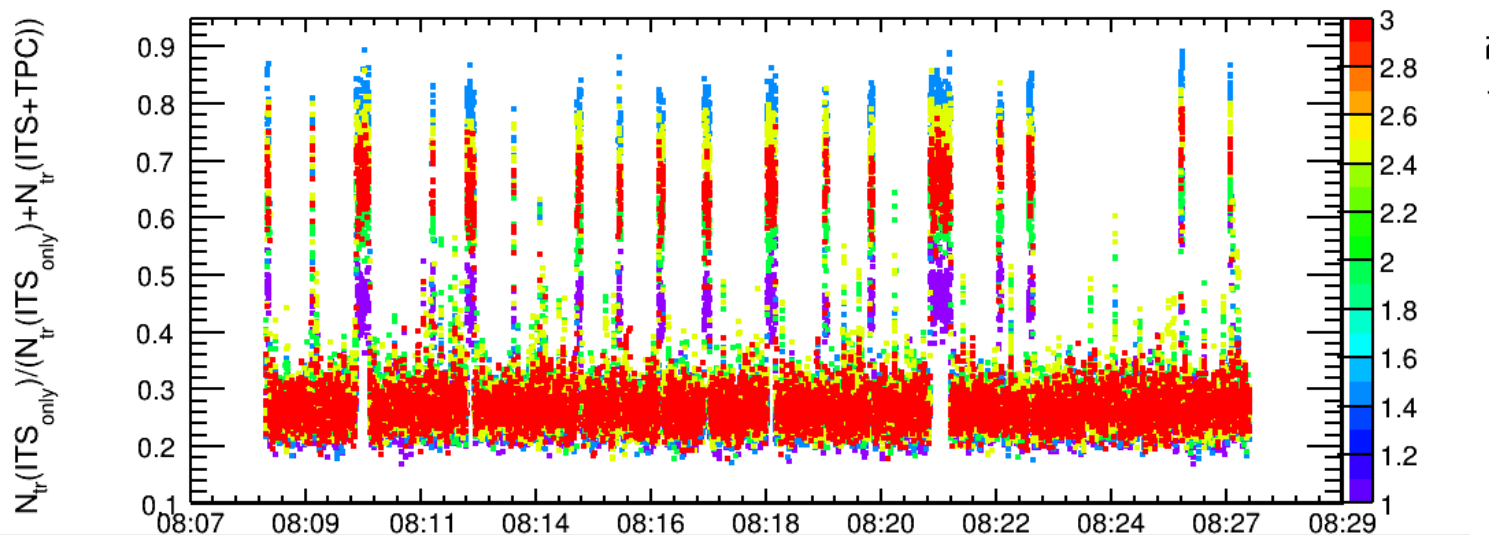
QA time series QA O(0.1s) - work in progress

Regular structure observed at sector boundary 2 in the run (**246272** and also some others)

Outliers in matching efficiency related to time intervals

- Is it OK?
- Which analysis is affected?
- What is the cause of outliers?
- Test "usual suspects" currently O(15 variables)

```
h+="outlierITSOnlyAC:Iteration$:timestampG:#((primMult>50&&abs(spdvz)<8)*phiCountAITSonly0)>>hisTimephiCountAITSonly0(2,-0.5,1.5,36,0,36,%0.0f,%f,%f)
",Int_t(timeEnd-timeStart)*10., timeStart,timeEnd);
h+="outlierITSOnlyAC:Iteration$:timestampG:#((primMult>50&&abs(spdvz)<8)*phiCountAITS0)>>hisTimephiCountAITS0(2,-0.5,1.5,36,0,36,%0.0f,%f,%f);"
,Int_t(timeEnd-timeStart)*10., timeStart,timeEnd);
```

Sum$(trdlmon.fElements):time {ir>0}
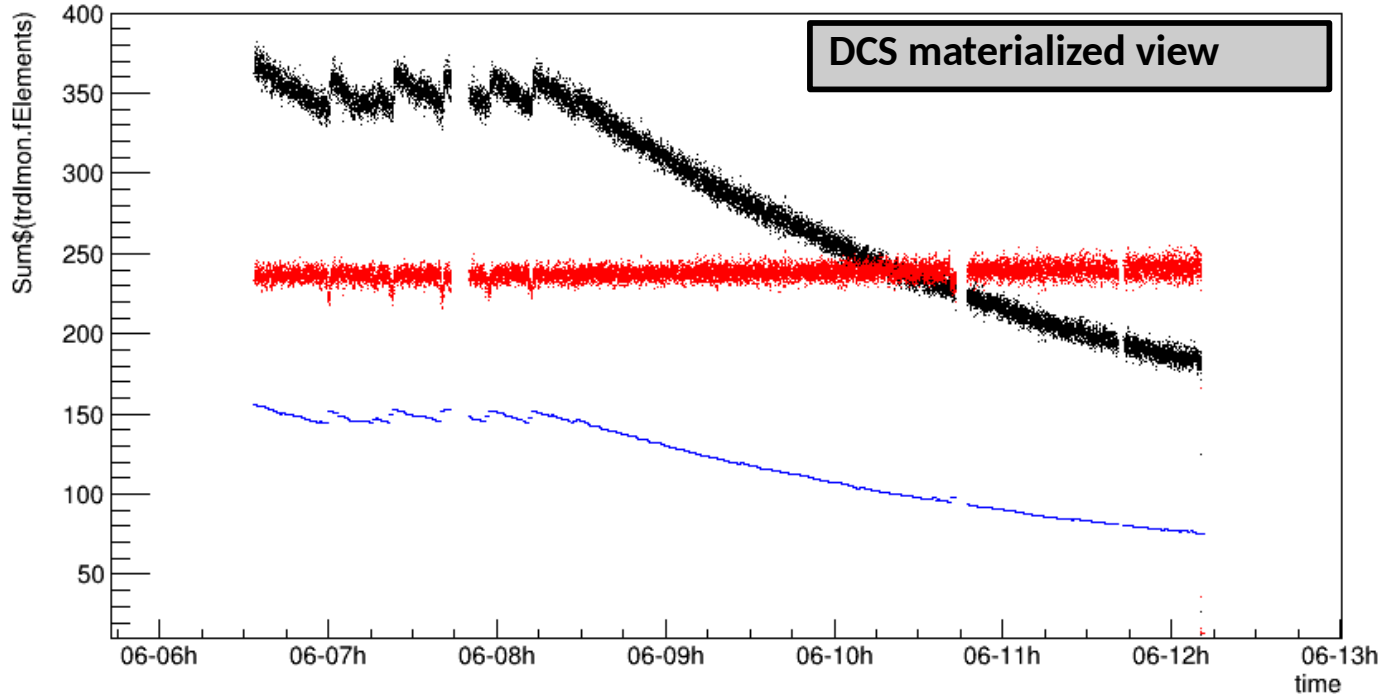
Are harware problems at sector 2 run **246272** related to some other harware changes (flux?)

No indication of flux modulation with the granularity O(min)

Flux estimators (a.u) in fill with run **246272** - limited time granularity >> 1 s

- **Sum of TRD currents, interaction rate, ratio Current/IR**
- IR levelling in fill of interest ~ 15 minutes
- TRD currents and IR estimator from trigger detector agree (no indication of background)

Classification problem:

- Anomalies/Outliers in the performance QA
- Find most relevant features in the other observables
  - Maps: currents, distortion, local multiplicity, matching efficiency, chi2, Ncl, resolution ....
  - Derived "invariant" variables:
    - e.g RMS of current map/phi averaged map/scaled maps
    - distortion map/phi averaged map
    - local discontinuities in time and space
    - "physics" acceptable performance

Explain/find hardware origin of anomalies (currently hierarchy of alarms)

Training data:

- current maps, and distortion maps (at GSI) can be exported to alien
- time bin based QA currently available only for few run

DCAr alarms decomposition

dEdx alarms decomposition

Proposal and conclusion

Pipeline with performance maps in N dimensions in form of generic function (TFormula).

- In many cases corresponding physical model or parameterization available

Consideration:

- Detector/Calibration/reconstruction expert
  - understand detector physics   easy to compare data with physical model or physical parameterization
- Time to access the information
  - Close to 0
- Information availability
  - Easy to share with others :
  - e-mail code snippet
  - JUPYTER notebook
  - interactive queries
- Algorithm standardization

$$f(p_0, p_1, p_2, \dots) \neq f_0(p_0) \oplus f_1(p_1) \oplus f_2(p_2) \oplus \dots.$$

Pipeline with performance maps in N dimensions in form of generic function (TFormula).

- In many cases corresponding physical model or parameterization available

Usage:

- differential QA
- understand/remember detector behavior - physical models
- find the detector/reconstruction/calibration limits/compare with intrinsic resolution
- predict/extrapolate performance modifying condition (e.g RUN3 IR → 50 kHz)
    - what happen if (common mode restoration algorithm fail)
- compare differentially data with MC
- make systematic error studies
- provide recipes for optimal cut selections
- enable tune on data in N-dimensions
- provide input/parameterizations for toy MC/fast MC
- feasibility studies
- **enable ML algorithms (tune on data)**

$$f(p_0, p_1, p_2, \ldots) \neq f_0(p_0) \oplus f_1(p_1) \oplus f_2(p_2) \oplus \ldots$$

Currently used N-dimensional analysis pipeline - ROOT5 based a working prototype

- Already in use for physics analysis

Pipeline presented to ROOT team (Axel)

Series of following discussion

- CSS, web server (Sergei)
- Visualization (Olivier)
- Queries, TDataFrame (Danillo)
- Database - query with TTree(SQL) m:n relationships, trees with incompatible (time/space) granularity (Jakob)
- Machine learning - TMVA, TFormula interface (Lorenzo)

Goal - (partial) integration with ROOT6 before ROOT workshop 2018 (September)

- current software back compatible - I use also ROOT6
- Soon to break forward compatibility

Visualization/Metadata/Webserver

- CSS → ROOT ? (ongoing)
- Extended visualization → JSROOT support for browsing
    - negotiating within ALICE group
- Web-server (restful API) resp. client queries (Boris)
- Documentation (ongoing discussion with Sergei, Olivier)

Just-in-time (JIT) compiler/Parallelization

- Code developed for ROOT 5, Root6/C++11 brings new possibilities
- Merge/Use TDataFrame functionality (in collaboration with Danillo)

Minimization/Machine learning regression

- More unit tests, set of predefined function/fitters, minimization methods
- TFormula interface to ML regression algorithms (TMVA::Reader) - agreed with Lorenzo
- Documentation

Local functional representation - local ND regression (THn interpolation ...)

- To be optimized (compact, smooth,fast, parallel)

Materialized view

M/N queries, aggregation

https://indico.cern.ch/event/611563/contributions/2607723/attachments/1474425/2283059/ROOT_IO_Workshop_TDF_170612.pdf

Examples

Recent usage of ND pipeline

- alarms example
- interactive n-dimensional browsing
- global/local fit example
- derived variable example
- consistency checks

- Space charge distortion calibration and studies

  - examples slides 7, 8

  - Distortion correlation studies

  - Symmetry plane fit demo

- Tracking performance characterization (TRD in combined tracking)

  - https://indico.cern.ch/event/710009/#53-trd-in-tracking-in-run2

- TPC/TRD tracking optimization

  - under development

- New event and track selection cuts (dEdx and chi2 QA analysis)

  - https://indico.cern.ch/event/710009/#62-proposal-for-new-event-and
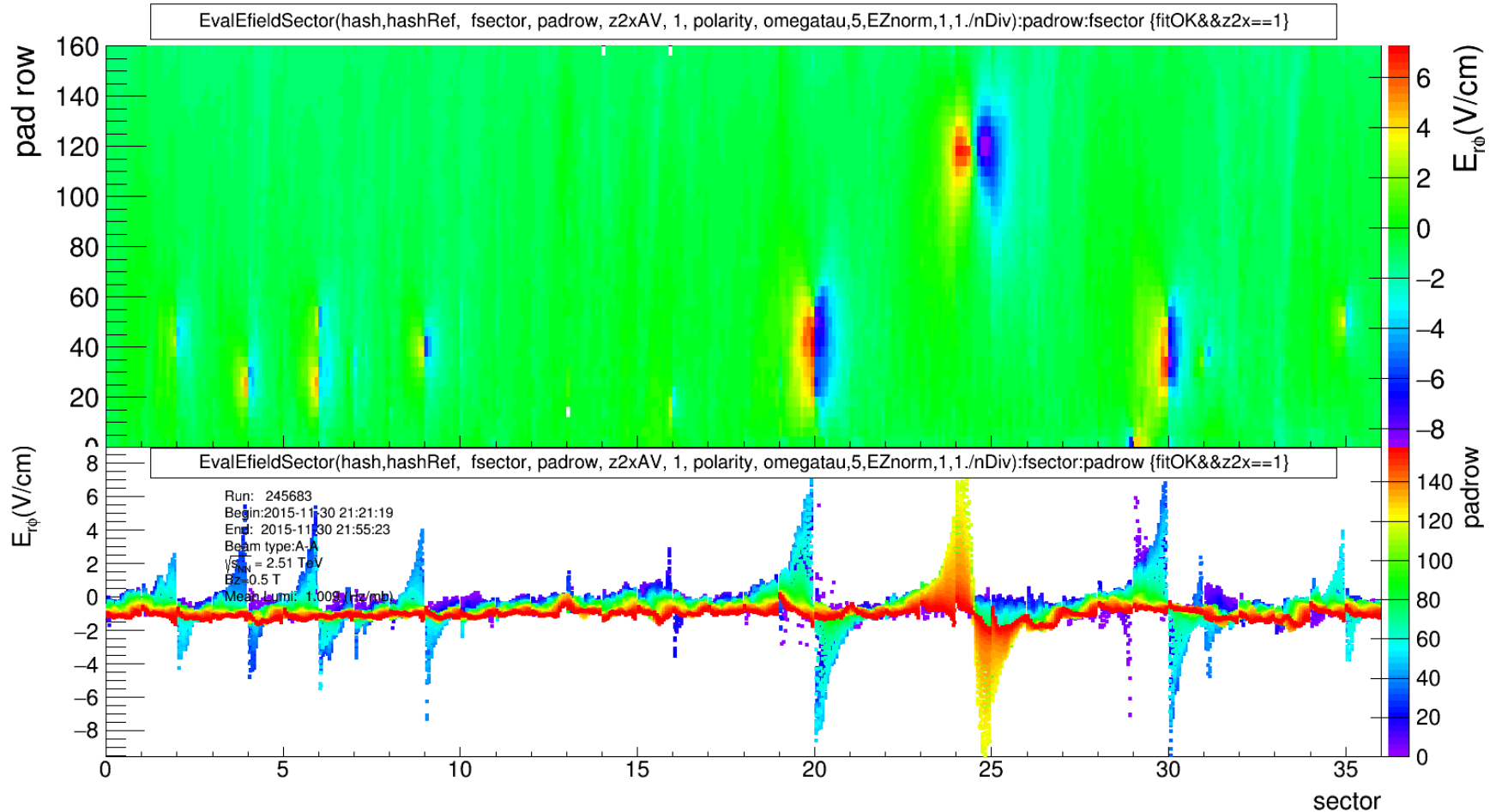
Run2:

Static view/projections without nominal map subtraction insensitive to the distortions

Run 3:

Width of the 1D projection distribution 20 cm. 1000 wider than requested resolution

Automatic alarms?
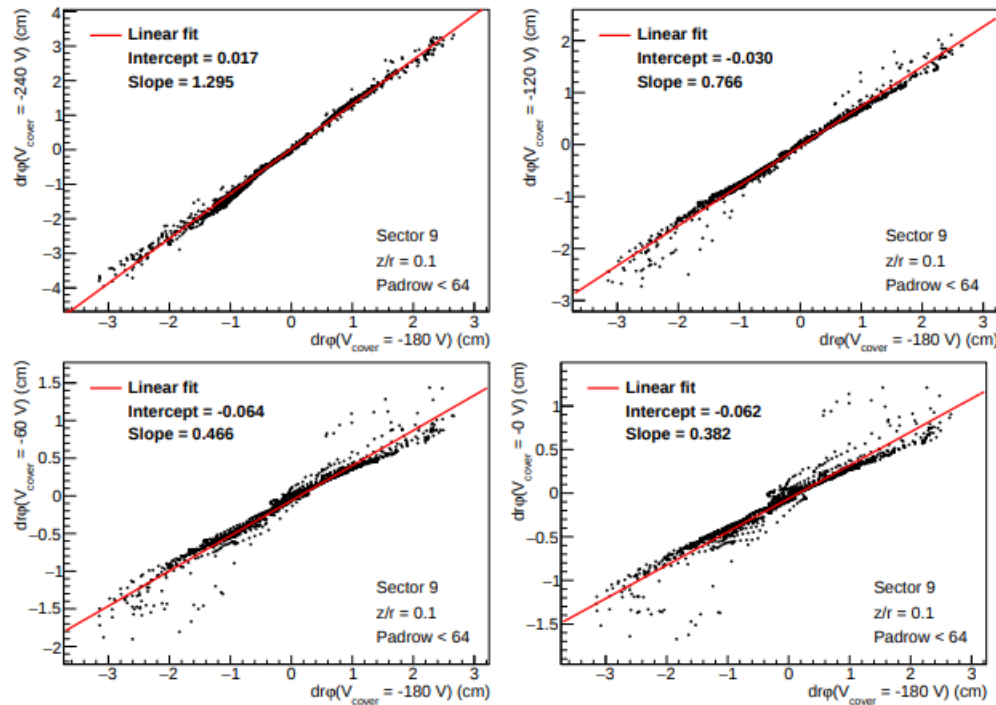
Figure 42: The four plots show the correlation between the distortions $dr\varphi$ with the nominal cover electrode voltage of $V_{cover} = -180\,V$ and the distortions with cover voltages of $-240\,V$ (top left), $-120\,V$ (top right), $-60\,V$ (bottom left) and $0\,V$ (bottom right) at sector 9. Each graph is fitted with a polynomial of first order. The fit parameters are quoted in the legend. The slope directly indicates the size of the distortions with the given cover electrode voltage relative to the nominal settings.

Alarms to be based on invariants - e.g. difference between maps and scaled reference maps

- Example: distortion maps correlation for different voltage setting
- Alarm criteria adjusted using models (sigmas) resp. history of past data
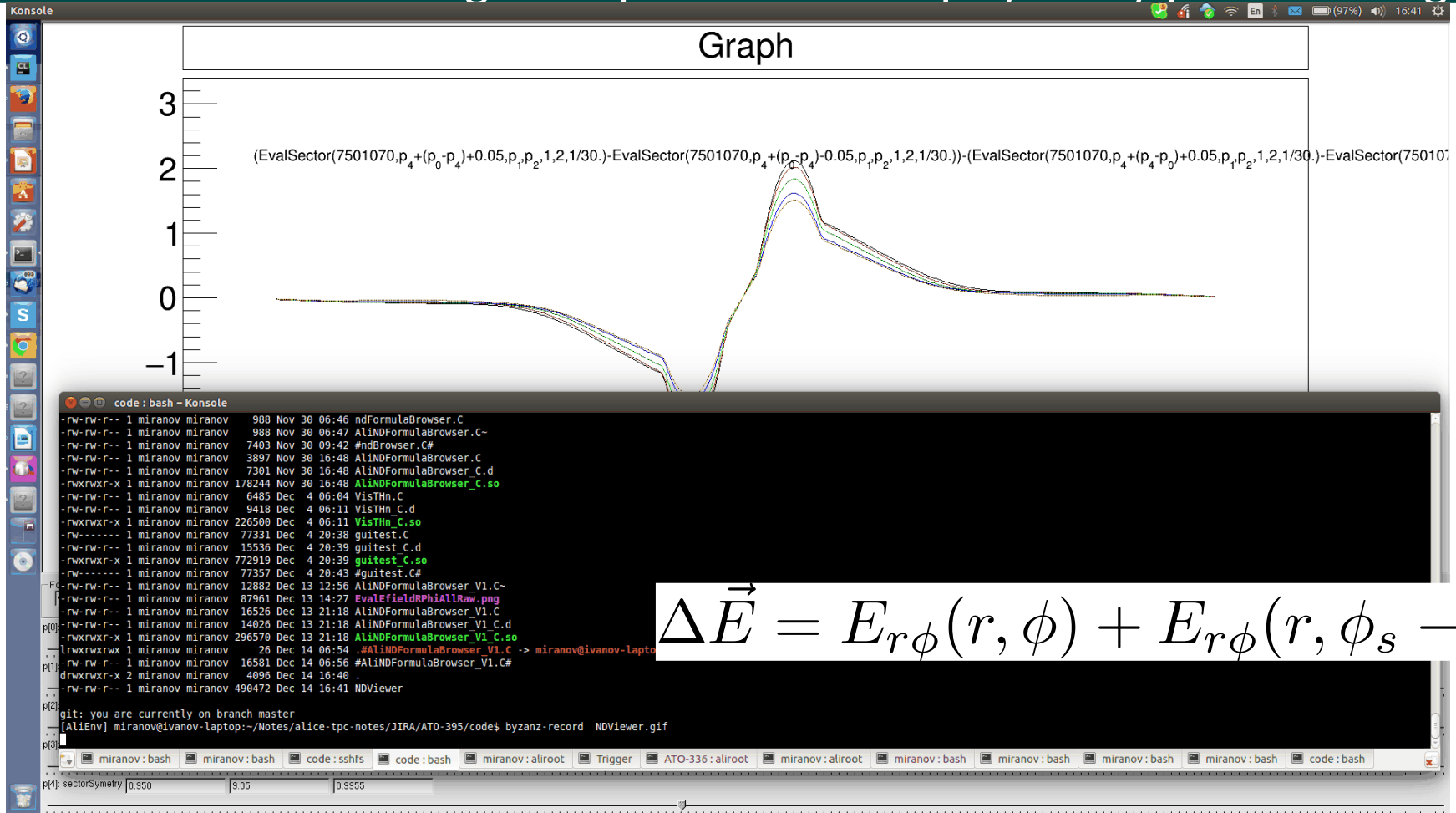
$$\Delta \vec{E} = E_{r\phi}(r,\phi) + E_{r\phi}(r,\phi_s - \phi)$$

Interactive finding asymmetry plane of the E field source:
- Set of sliders to navigate and visualize E field in 4D (mirror plane as dimension 4) - AliNDLocalRegression
- Example: map -mirrored map as used to find origin of distortion

$$p(I|(i - \Delta||i + \Delta) = p_0 + p_1 \times (dEdx \times L)^{p_2} (p_4 + occu^{p_3})$$

Model:

```
AliTMinuitToolkit *fitterNclFraction = new AliTMinuitToolkit("AliTMinuitToolkitFitterNclFraction.root");
TFormula *formulaNclFraction = new TFormula("formulaNclFraction", "[0]+[1]*((x[0]^[2])*([4]+x[1]^[3]))");
fitterNclFraction->SetFitFunction((TF1 *) formulaNclFraction, kTRUE);
```

Fit:

```
fitterNclFraction->FillFitter(treeMap, (fitValue + ":1/0.01").Data(), "(50./((sdEdxMaxCenter**2)*sqrt(1+atglCenter**2))):(multACenter/10000)", "fitCutNcl&&Entry$%3==1", 0
                              10000000);
fitterNclFraction->Bootstrap(20, "report1");
```

Visualization/query:

```
treeMap->SetAlias((fitValue + "FitM").Data(), fitterNclFraction-
>GetFitFunctionAsAlias().Data());
treeMap->Draw("missingFrac0FitM:mdEdx:multACenter1000", "fitCutNcl&&atglBin==1", "colz");
TStatToolkit::AdaptHistoMetadata(treeMap, 0, "colz");
```

Number of clusters along the trajectory determined by track length

The local probability cluster is assigned to track

p(I) Conditional probability cluster exists in layer I a cluster exists in local neighborhood:

- cluster below pedestal+threshold
- cluster rejection because of chi2 distance (position bias due to local overlap)
- fit in 3D (dEdx, L, occu) with 5 parameters (AliTMinuttoolkit+Bootstrap)

Pedestal at high multiplicity biased and fluctuating:

- In the first approximation pedestal bias is proportional to the occupancy $p_3 \sim 1$
- Probability signal above threshold ~ 1/(L x dEdx) $p_2 \sim -1$

Rough linear scaling with occupancy and 1/dEdx (gain) and tan(Θ)

Global fit describes the data in wide range of multiplicities, dEdx and angles

Extrapolation to RUN3

- ~45000 - 40-50 % loss ?
- algorithm to deal with missing clusters

```
makeNDLocalFit("hisRatioMax01Dist.rmsG","isOK","sqrt(0.01**2+hisRatioMax01Dist.rmsG**2/hisRatioMax01Dist.entries)");
makeNDLocalFit("hisRatioMax12Dist.rmsG","isOK","sqrt(0.01**2+hisRatioMax12Dist.rmsG**2/hisRatioMax12Dist.entries)");
makeNDLocalFit("hisRatioMax02Dist.rmsG","isOK","sqrt(0.01**2+hisRatioMax02Dist.rmsG**2/hisRatioMax02Dist.entries)");
treeMap->SetAlias("rmsMax0","sqrt((hisRatioMax01Dist.rmsGFit**2+hisRatioMax02Dist.rmsGFit**2-
hisRatioMax12Dist.rmsGFit**2)/2.)");
```

$$RMS_{Qi} = \sqrt{RMS^2_{Qi/Qj} + RMS^2_{Qi/Qk} - RMS^2_{Qi/Qk}/2}$$

$$RMS_{ROC} \times \sqrt{N_{CR}} \approx p_0 \left(dEdx^{p_1} \times \sqrt{(1 + \tan(\theta))^2}^{p_2} \right)$$

$$(1)$$

Relative intrinsic resolution estimated from the RMS of the dEdx ratios histograms

At low multiplicity relative dEdx resolution - power law scaling with the dEdx and tracklet length

Parameters p0, p1 and p2 fitted in global fits
- fit using AliTMinuitToolkit + bootstrap method
- error estimator as obtained in bootstrap

rmsMax0NormFit0:sdEdxMaxCenter:atglCenter {fitCut0&&multABin==1}

rmsMax0NormLF:rmsMax0NormFit0:atglCenter {fitCut0&&multABin==1}

$$RMS_{Qi} = \sqrt{RMS^2_{Qi/Qj} + RMS^2_{Qi/Qk} - RMS^2_{Qi/Qk}/2}$$

$$(1)$$
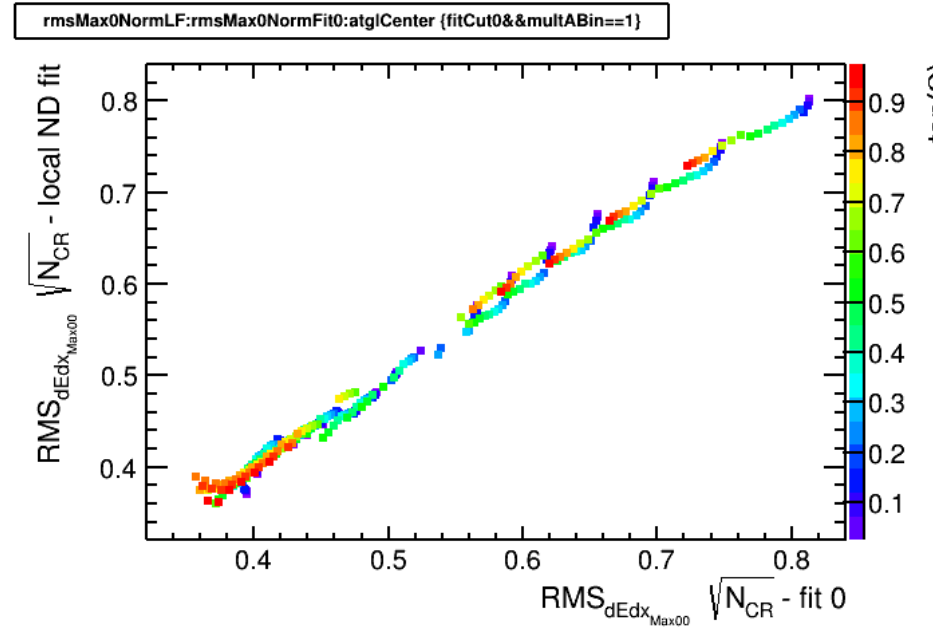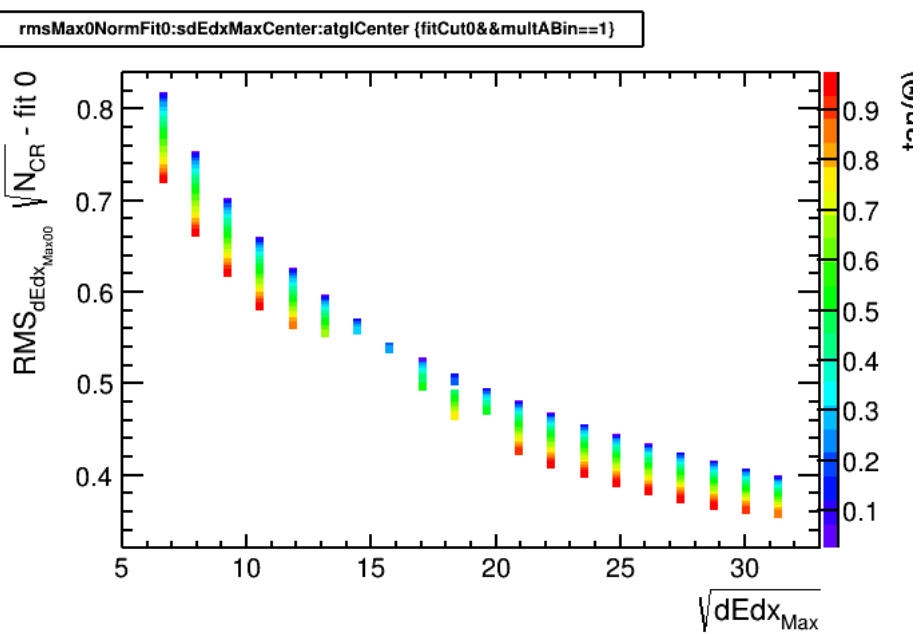
$$RMS_{ROC} \times \sqrt{N_{CR}} \approx p_0 \left(dEdx^{p_1} \times \sqrt{(1 + \tan(\theta))^2}^{p_2} \right)$$

Relative Intrinsic resolution for IROC as parameterized by global resp. local fit

At low IR agreement between dEdx resolution and power low model

- p0 =1.96 +-0.01
- p1=0.233+-0.001
- p2=0.40+-0.02

```
void CheckInvariant(){

  treeMap->Draw("(exp(hisRatioMax01Dist.meanGFit)/exp(hisRatioTot01Dist.meanGFit))/
(hisRatioTotMax1Dist.meanGFit/hisRatioTotMax0Dist.meanGFit)","isOK&&entries>200");
  treeMap->Draw("(exp(hisRatioMax02Dist.meanGFit)/exp(hisRatioTot02Dist.meanGFit))/
(hisRatioTotMax2Dist.meanGFit/hisRatioTotMax0Dist.meanGFit)","isOK");
  treeMap->Draw("(exp(hisRatioMax12Dist.meanGFit)/exp(hisRatioTot12Dist.meanGFit))/
(hisRatioTotMax2Dist.meanGFit/hisRatioTotMax1Dist.meanGFit)","isOK");
}
```

$$Q_{\det_0,\text{type}_0}/Q_{\det_1,\text{type}_1} \neq const$$

Non invariant

$$Q_{\det_0,\text{type}_0}/Q_{\det_1,\text{type}_1} \approx \frac{dEdx \times T_{\det_0,\text{type}_0}}{dEdx \times T_{\det_1,\text{type}_1}} \qquad (1)$$

$$T_{\det_i,\text{type}_j} \approx f(dEdx, \tan(\theta), \tan(), Gain, occu, thr.)$$

$$\frac{\dfrac{<Q_{max0/max1}>}{<Q_{to0/tot1}>}}{\dfrac{<Q_{tot0/max0}>}{<Q_{tot1/max1}>}} = const$$

Performance map invariant

Set of 3-4D transfer function/maps created

Function to be used in user analysis (track and event cuts, tune on data)

Internal consistency of maps can be checked using invariants

New use cases

Set of 3D distortion maps to be taken with O(s) granularity

Set of 3D digital (O(500000) pixels and analog currents maps to be taken with O(ms) time granularity

To be QA-ed with appropriate time granularity

- in case of problems later will be not recoverable
- decision/correction to be made online
- outliers will be always present
  - long tails - asymmetric distribution of number of pile-up tracks, ion cloud

Huge spread of distortion O(20 cm) - position and time (gain, epsilon, local density) dependent

Alarms to be based on the set of "invariants"

- Local smoothness in space and/or time
  - Smoothness assumption in time
    - steps in the IR are normal/ some double ratios to be used
    - left/right median filter
- Variable/local median
  - alarms on the mean/median/rms/cumulant of distributions

RUN3:
more complex than Run2 - good diagnostic needed
many decisions to be made on-line. In case lossy compression used, we can not afford mistakes

PbPb 2015 - time bin QA time series
Run2 example of ongoing investigation

ntrITSRatio:T:sectorBin {entries>50&&(sectorBin==2||sectorBin==4||sectorBin==6||sectorBin==10)}
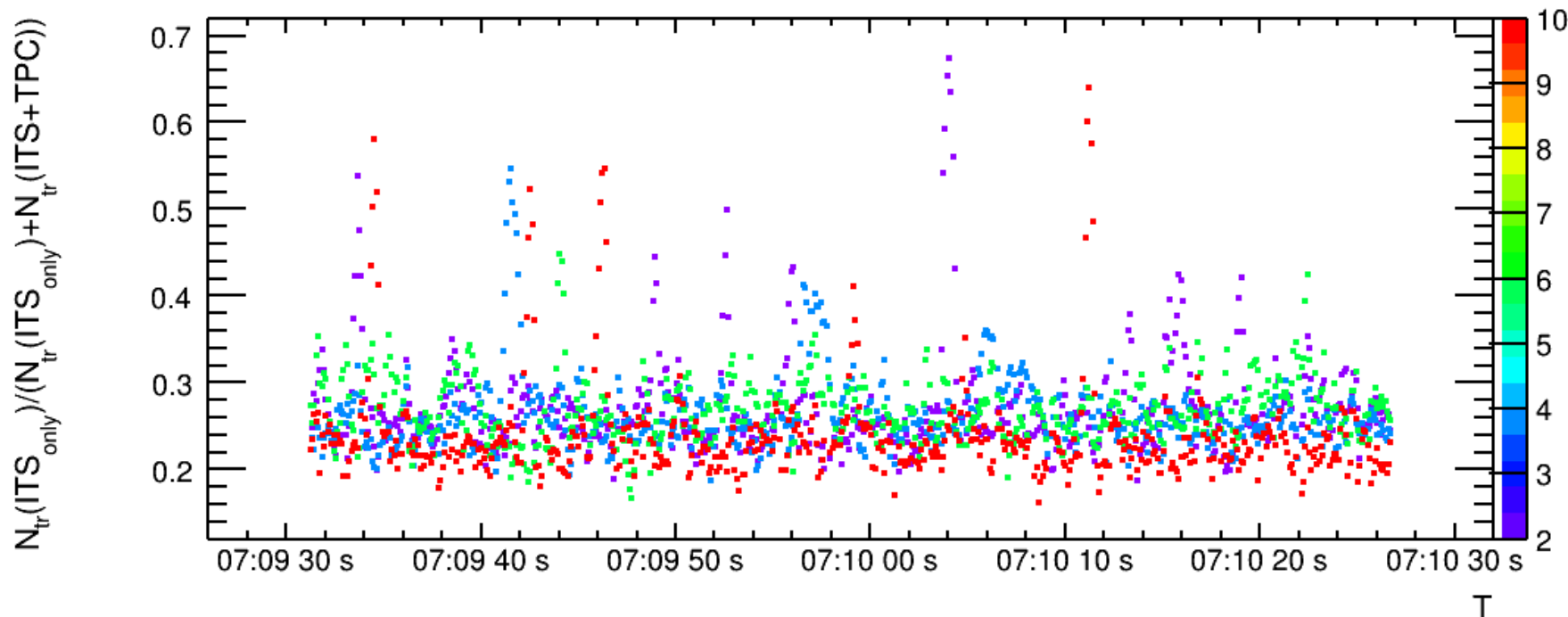
Time series QA O(0.1s)

In particular time intervals **O(0.5) seconds** distortion increased
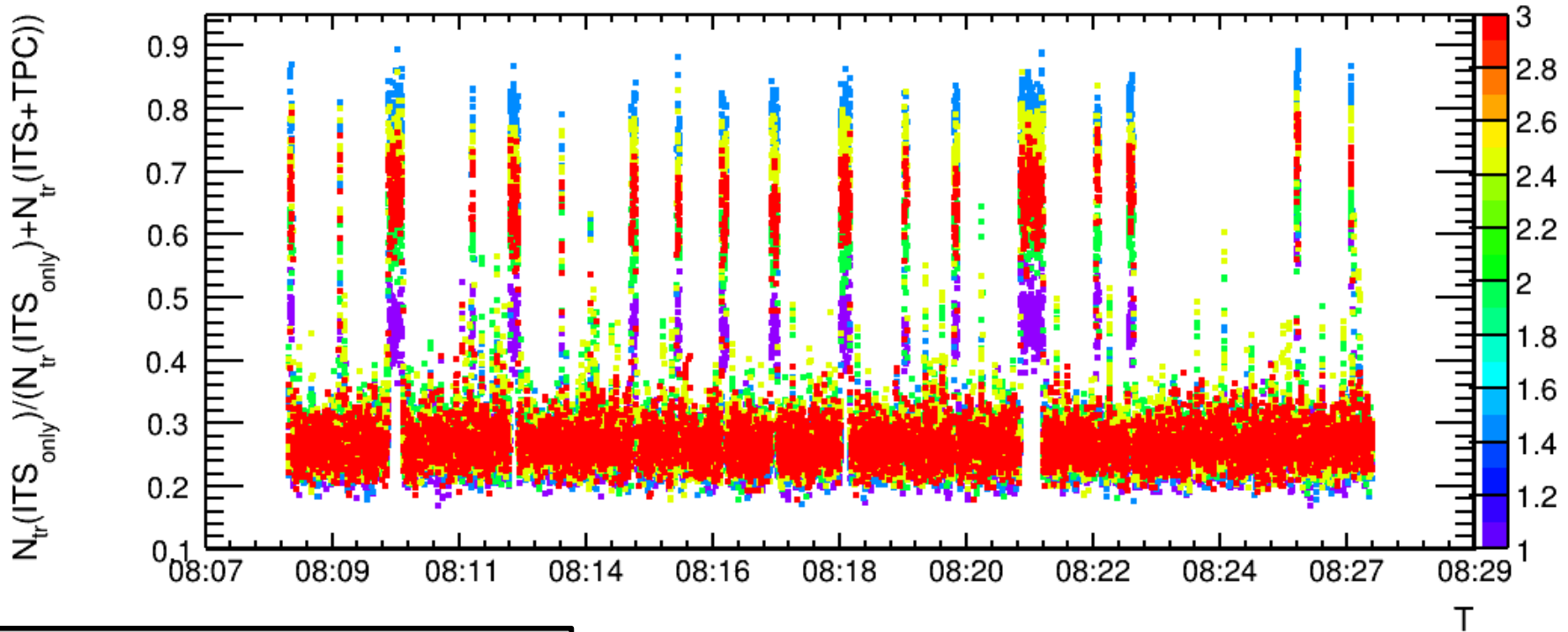
- locally worse resolution and matching efficiency

Distortion independent- see time position of spikes

- **sector bins 2, 4, 6, 10**

Correlated TPC/ITS efficiency loss. Problematic run 246272 begin

ntrITSRatio:T:sectorBin {entries>50&&abs(sectorBin-2)<1.1}

Time series QA O(0.1s)

Regular structure observed at sector boundary 2 in the run (246272 and also some others)

Outliers in matching efficiency related to time intervals

- Looks like regular position O(min) spacing
- Irregular amplitude, probability and duration
  - begin of run (higher IR) - bigger probability longer duration
- Possible origin - "onset of distortion"
  - 2 line charges holes in sector boundary 2 - one big one small

Sum$(trdlmon.fElements):time {ir>0}

DCS materialized view
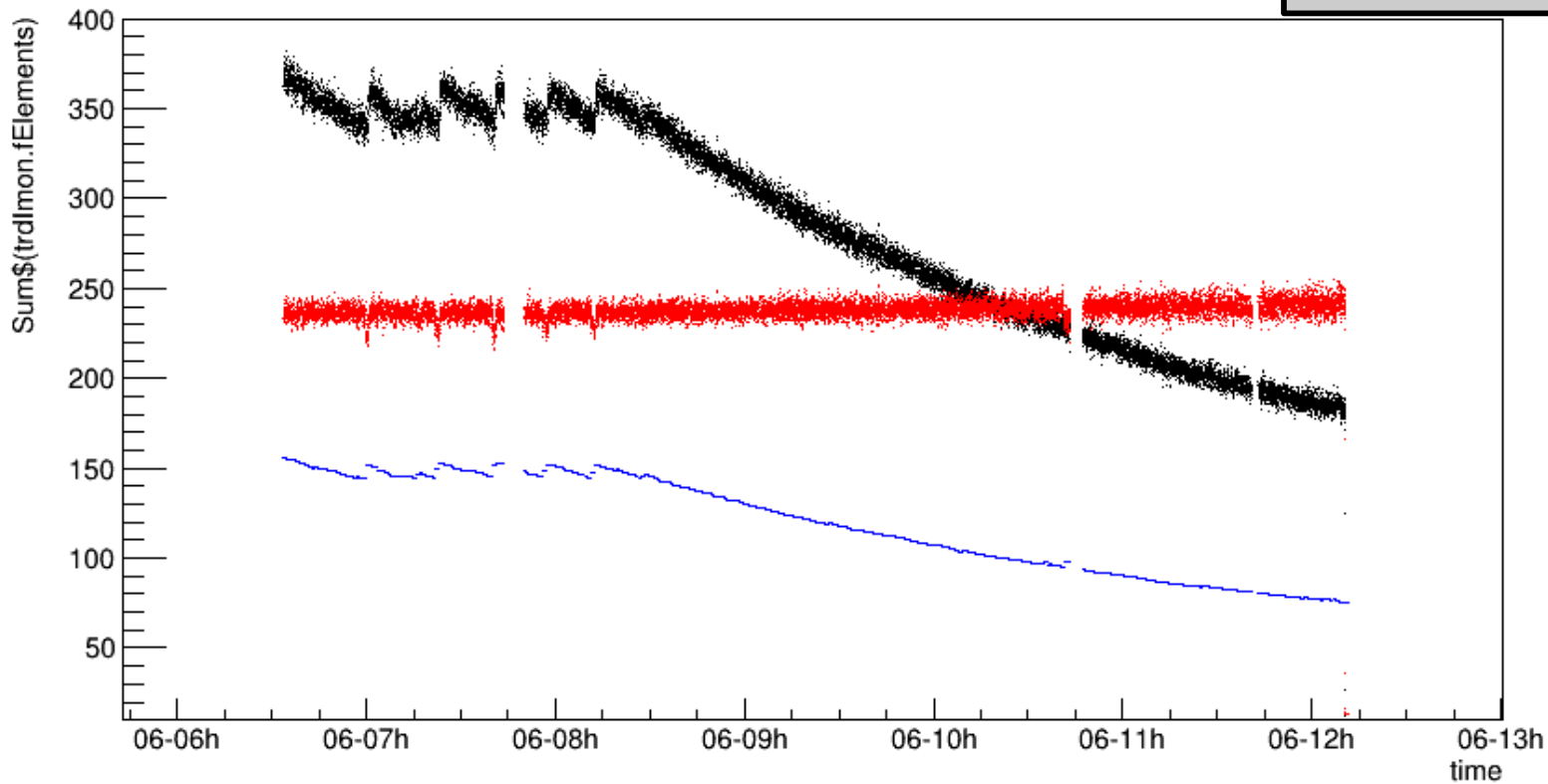
No indication of flux modulation with the granularity O(min)

Flux estimators (a.u) in fill with run 246272 - limited time granularity >> 1 s

- **Sum of TRD currents, interaction rate, ratio Current/IR**
- IR levelling in fill of interest ~ 15 minutes
- TRD currents and IR estimator from trigger detector agree (no indication of background)

valueAnodeMedLayer.fElements:time {abs(valueAnodeMedLayer.fElements-1)<1 && vecLayer.fElements==2}
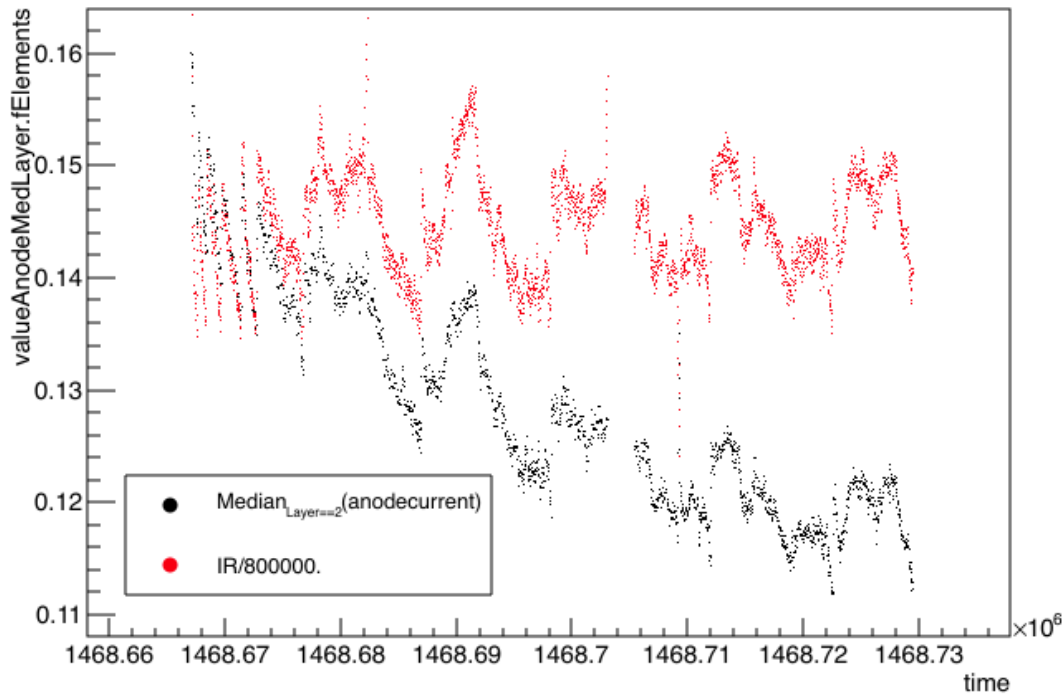
spikes in hardware currents and IR
not all time intervals usable
Model validity +-10 %

**DCS materialized view
TRD,GRP**

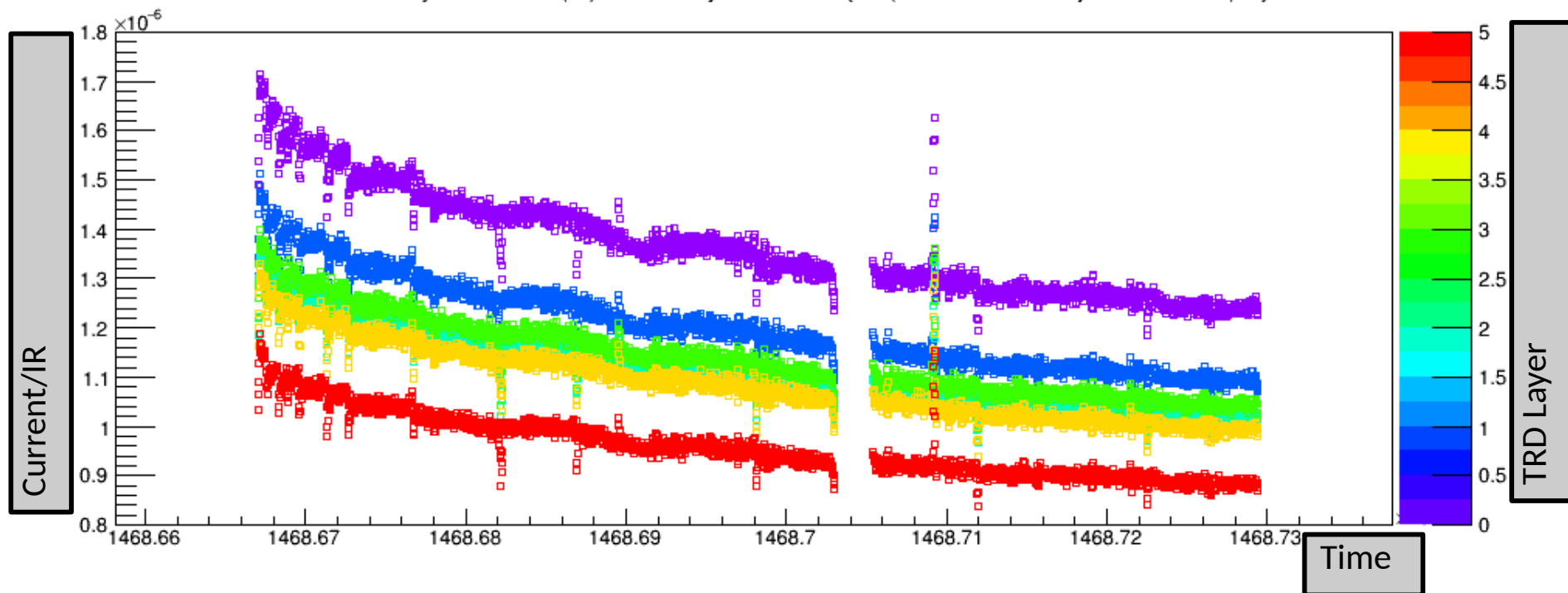**TPC distortion corrected using the IR resp. current estimator**

- applicability of distortion scaling models limited to narrow ranges of possible values
- High time gradient → biased estimators

Better estimator to be obtained using TRD currents

- TRD currents ~ signal + beam background
- Weighting of currents → distortion scaling. Appropriate extrapolation function to TPC to be selected and controlled

valueAnodeMedLayer.fElements/(IR):time:vecLayer.fElements {abs(valueAnodeMedLayer.fElements-1)<1}

TPC distortion corrected using the IR estimator

**Better estimator to be obtained using TRD currents**

- Current normalized to IR from scalers in individual layers (color decreasing as function of R)

- decrease as expected - background is decreasing IR is leveled

- Steps at region with high gradient of IR - 30 %

    - **Time intervals to be removed from analysis?** - time series analysis needed to decide

matchEffITS:nITSOnlyARel {primMult>100&&trdMult>200&&nITSOnlyARel<2&&abs(spdvz)<8}
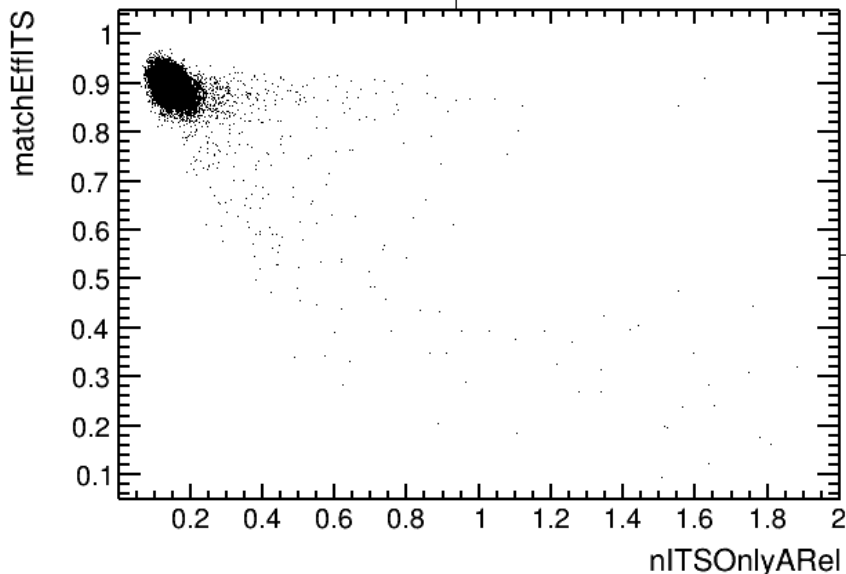
nITSOnlyARel:primMult/tpcMult {primMult>100&&trdMult>200&&nITSOnlyARel<2&&abs(spdvz)<8}
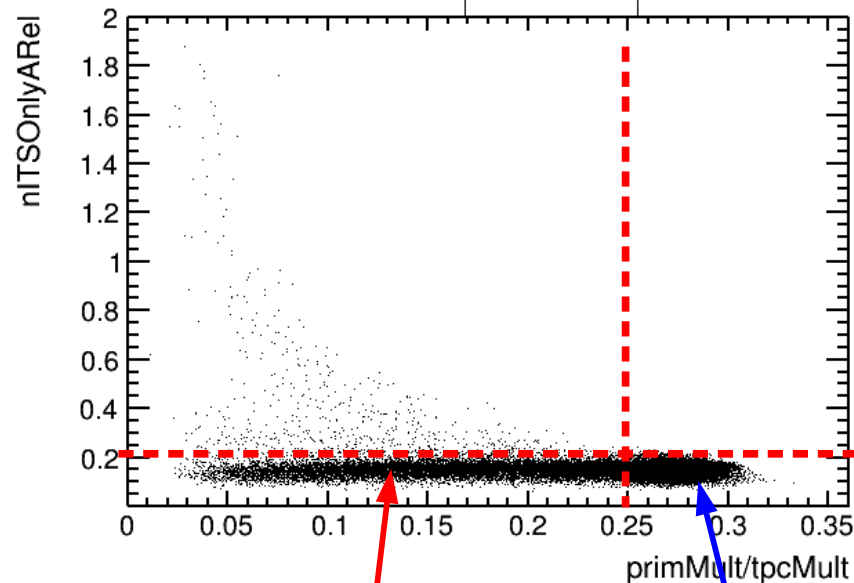
Event+TPC pileup   Normal event

TPC → ITS conditional matching efficiency related to the probability to find complementary ITS tracks

Median value of the fraction of complementary tracks does not depends strongly on the pile-up admixture

- Outliers in efficiency for small fraction of high pile-up events

Hypothesy:

- **TPC pileup events pointing to the ITS (small fraction) fights for the ITS clusters**
- **Correlated degradation of the performance**
    - degradation proportional to relative difference of mutliplicities

gainMIP:time:occMean {abs(dits)<3600&&run>244000&&abs(gainMIP_medianLeft-gainMIP_medianRight)<0.05}
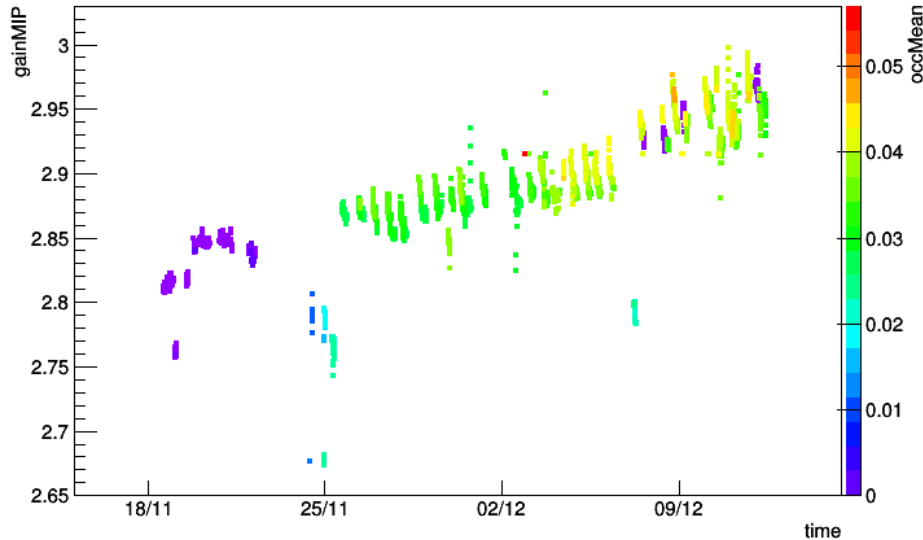
gainMIP/gainMIP_median-1:time:occMean {abs(dits)<3600&&run>244000&&abs(gainMIP_medianLeft-gainMIP_medianRight)<0.05}

G-<G>
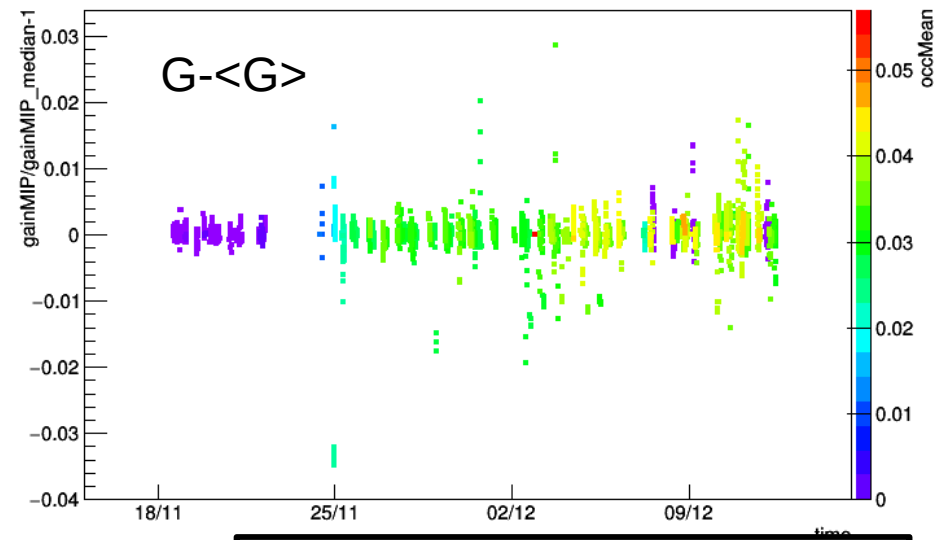
**DCS materialized view
TRD,GRP
Aggregated information query
example**

## Left: Gain calibration as function of time (O(min)) time intervals

- Interaction rate modulation due imperfect ion tail/common mode correction (without Iontail/common mode correction O(20-30%) decrease of Q per central event ())

- Pressure and temperature corrected automatically, using calibrated curve. Gain calibration to account for change of gas composition

Right:

- Gain - local median gain (+-20 min). Outlier time intervals O(+-2%) visible

- Could be problematic for correlation studies

nITSOnlyARel>0.25:primMult:tpcMult {primMult>200&&abs(spdvz)<8}

primary multiplicity

efficiency outlier fraction

tpc multiplicity
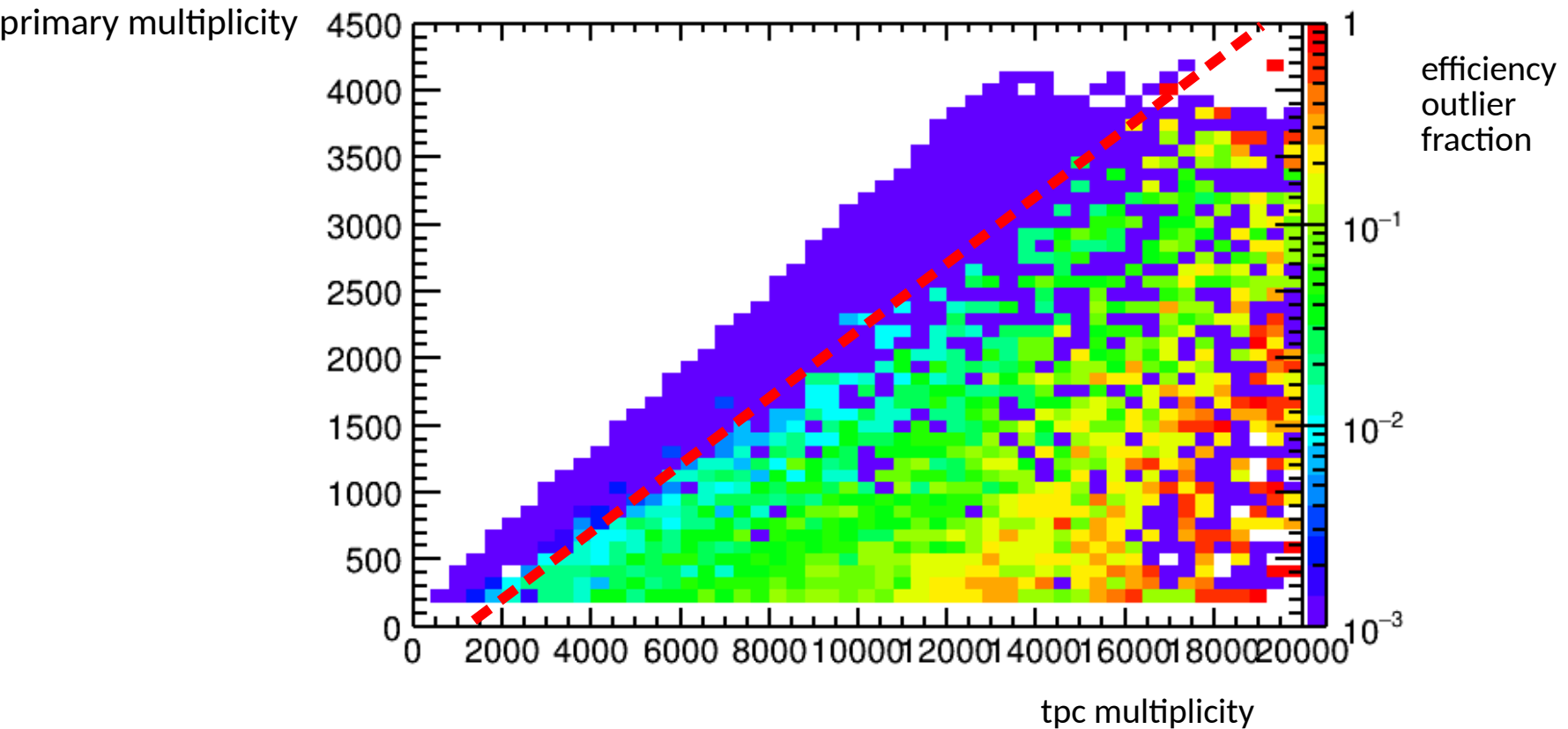
In current analysis of the PbPb LHC15o strong pile-up rejection cut used

Suggestion - only very small part of pile-up events are outliers in matching efficiency

Better tagging of outliers using mean event properties
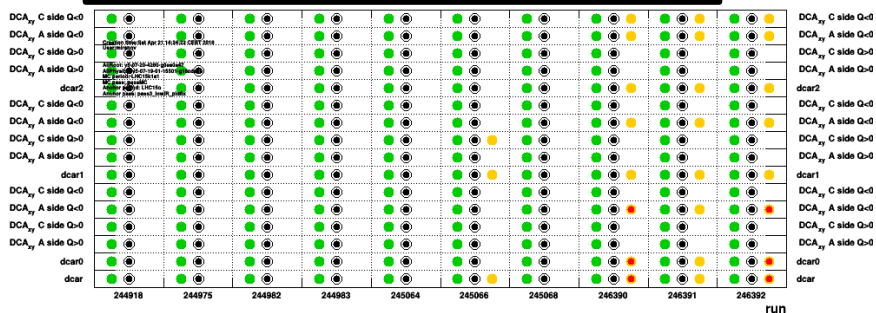
Classification problem:

- Anomalies/Outliers in the performance QA
- Find most relevant features in the other observables
  - Maps: currents, distortion, local multiplicity, matching efficiency, chi2, Ncl, resolution ….
  - Derived "invariant" variables:
    - e.g RMS of current map/phi averaged map/scaled maps
    - distortion map/phi averaged map
    - local discontinuities in time and space
    - "physics" acceptable performance

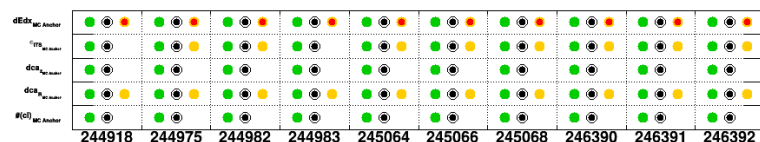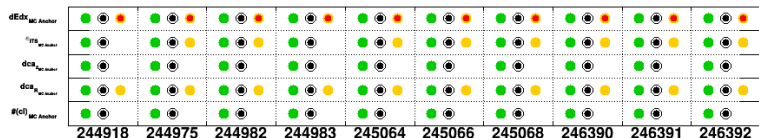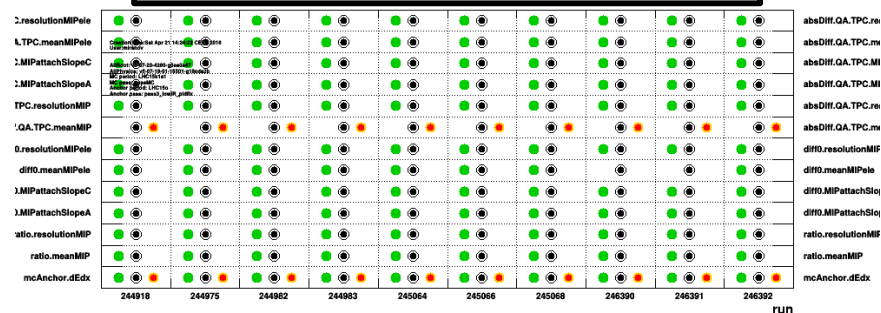## Explain/find hardware origin of anomalies (currently hierarchy of alarms)

Training data:

- current maps, and distortion maps (at GSI) can be exported to alien
- time bin based QA currently available only for few run

DCAr alarms decomposition

dEdx alarms decomposition

Input:
Kinematics (particle properties at vertex)
Track references (particle properties along trajectory)

Toy MC
Toy reconstruction

Output:
Ideal acceptance maps
Ideal efficiency maps
Ideal Matching efficiency maps
Ideal performance/resolution maps

Spectra/Maps in bins:
tracks: pt, eta, PID
events: Multiplicity in eta bins 2-3D (~SPD, ~V0)
ET

Figure.

Ongoing - discussion with Anton, Markus

Precision of machine learning algorithm limited by knowledge of ground truth (in out case MC description)

- example analysis J/psi studies (ongoing Lukas Layer)

Features:

- chi2, missing clusters
- DCA
- PID probability in TPC, TOF
- multiplicity
- pt, eta

Precision of MC limited (data usually worse), not all effects simulated:

- pile-up, ion-tail, common mode
- distortion fluctuation
- unknown detector effect … (trying to understand)

Solution to tests:

- Remap/retune MC feature to behave as in real data using performance maps
  - e.g. using performance maps ratios, diffs, n-sigma ranges
  - operation depends on the variable of interest - e.g chi2 using ratios

$$f(p_0, p_1, p_2, \dots) \neq f_0(p_0) \oplus f_1(p_1) \oplus f_2(p_2) \oplus \dots.$$

Tracking performance not homogeneous in space

- discussed with Jet group (Leticia and Markus)

MC - using add hoc parameterization of the distortion fluctuation

- only partial description of performance deterioration
- physical model of fluctuation - no time to implement

Proposed solution:

- unfolding spectra using the 4-5D (phi eta performance maps)
- folding MC to emulate additional deterioration of performance
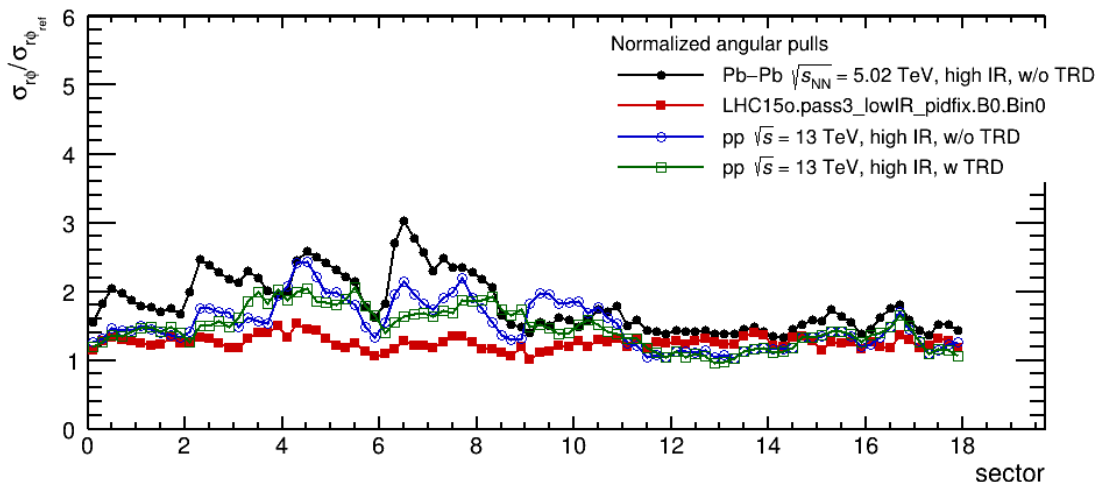
$$f(p_0, p_1, p_2, ...) \neq f_0(p_0) \oplus f_1(p_1) \oplus f_2(p_2) \oplus ....$$

Fast MC developed for RUN3 (Ruben, Johannes …) used for …

- fast particle transport - trajectories, interaction, energy loss MS
- fast reconstruction of the data

Detector response **parameterized using N dimensional maps/parametrization** (Johannes, Marian):

- Distortion maps
- Dead zone
- Cluster smearing
 - TPC-ITS matching efficiency (https://alice.its.cern.ch/jira/browse/AOC-13)

Using proper detector response parameterization more realistic performance than using current full MC expected

- Activities interrupted after Johannes left collaboration
- Reactivation (?) as one of the transport model

Several **other options** investigated in context of RUN3 activities. I consider N-Dimensional parameterization option as **more promising (either NDlocal regression or ML regression)**

- more exotic options (GANs) to be benchmarked in respect to "more" classical
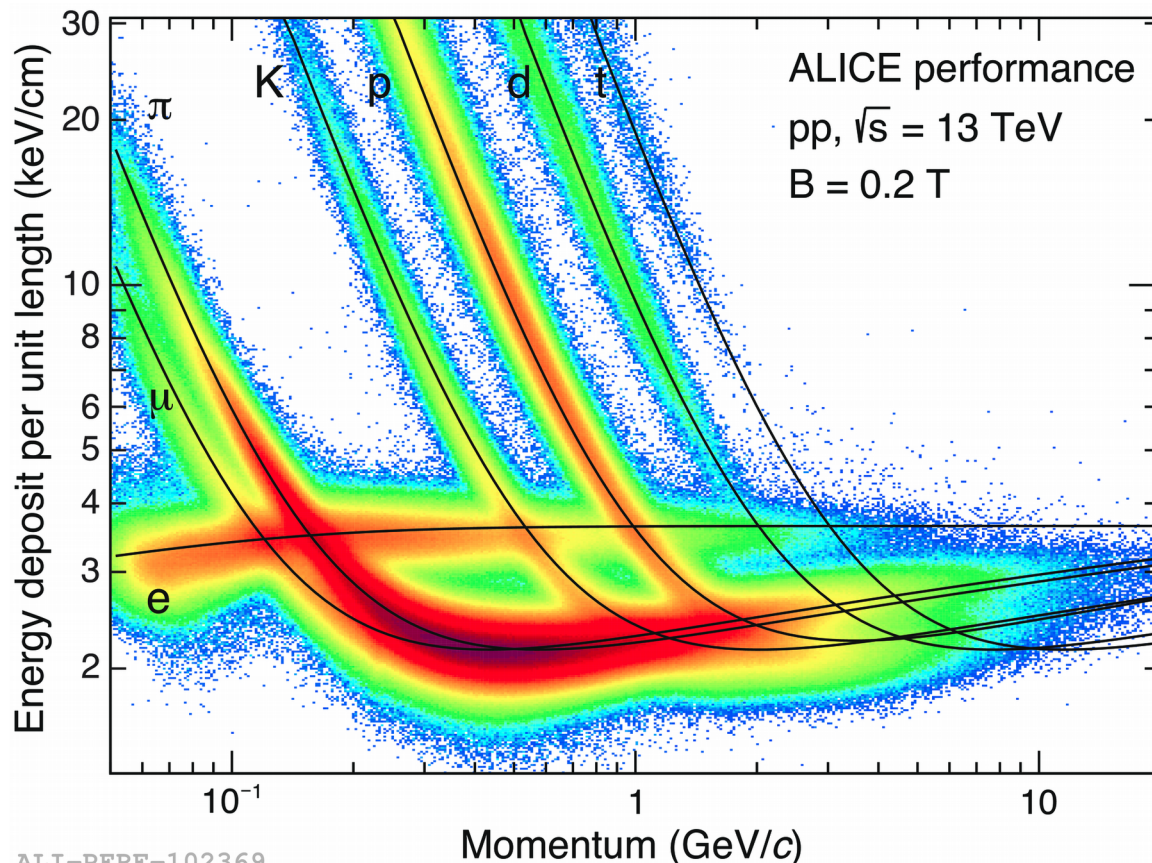
Non standard tracking and PID:

Tracking down to 30-40 MeV
TPC standalone without outer detectors
Pile-up rejection using dEdx invariant
First usage of transfer function
        Additional correction for the DCA
Additional dEdx correction for the gain loss and electron attachment
in particular run gain was 2 times smaller than nominal

https://aliceinfo.cern.ch/Figure/node/8731

C. Patrignani et al. (Particle Data Group), Chin. Phys. C, 40, 100001 (2016) and 2017 update.

TPC standalone tracking and dEdx performance down to 30-50 MeV (ATO-245)

Using **TPC dEdx invariant cut ("PID likelihood")** and 3D transfer function correction, TPC standalone analysis enabled

Usage/TO DO/Conclusion

Pipeline with performance maps in N dimensions in form of generic function (TFormula).

- In many cases corresponding physical model or parameterization available

Usage (/no black box approach):

- understand/remember detector behavior - physical models
- find the detector/reconstruction/calibration limits/compare with intrinsic resolution
- predict/extrapolate performance modifying condition (e.g RUN3 IR → 50 kHz)
  - what happen if (common mode restoration algorithm fail)
- compare differentially data with MC
- make systematic error studies
- provide recipes for optimal cut selections
- enable tune on data in N-dimensions
- provide input/parameterizations for toy MC/fast MC
- feasibility studies
- **enable ML algorithms (tune on data)**
- **N dimensional queries - web server/client**

$$f(p_0, p_1, p_2, \ldots) \neq f_0(p_0) \oplus f_1(p_1) \oplus f_2(p_2) \oplus \ldots.$$

Currently used N-dimensional analysis pipeline   - a working prototype

- Already in use for physics analysis

Visualization/Metadata

Integration with ROOT6 Functionality for ROOT6?

- CSS → ROOT ? (ongoing)
- Extended visualization → JSROOT support for browsing
  - negotiating within ALICE group
- Web-server (restful API) resp. client queries ()
- Documentation (ongoing discussion with Sergei, Olivier)

Just-in-time (JIT) compiler/Parallelization

- Code developed for ROOT 5, Root6/C++11 brings new possibilities
- Merge/Use TDataFrame functionality (in collaboration with Danillo)

Minimization/Machine learning regression

- More unit tests, set of predefined function/fitters, minimization methods
- TFormula interface to ML regression algorithms (TMVA::Reader) - agreed with Lorenzo
- Documentation

Local functional representation - local ND regression

- To be optimized (compact, smooth,fast, parallel)

Materialized view, M/N queries, aggregation

https://indico.cern.ch/event/611563/contributions/2607723/attachments/1474425/2283059/ROOT_IO_Workshop_TDF_170612.pdf

Multidimensional analysis using ND pipeline

Successfully used for the reconstruction and calibration (development/commissioning):

- Space charge distortion calibration and studies
- Tracking performance characterization (TRD in combined tracking)
- TRD tracking optimization
- New event and track selection cuts (dEdx and chi2 QA analysis)

Visualization development (see next talk - Boris)

Future plans usage of the ND pipeline:

- Optimization of the physics analysis (jets, J/psi, mass analysis using TOF)
- Tune on data and fast MC
- Run3, O2

All development using old ROOT5 and old C++ (currently used in AliRoot)

- Will strong profit from new C++, JIT compiler and new ROOT6 functionaltiy
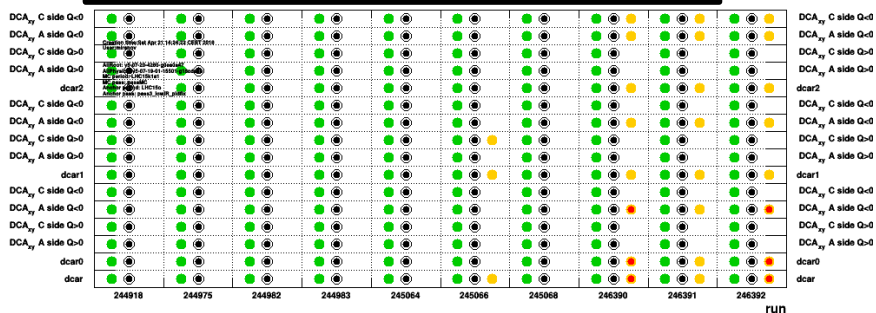
Classification problem:

- Anomalies/Outliers in the performance QA
- Find most relevant features in the other observables
    - Maps: currents, distortion, local multiplicity, matching efficiency, chi2, Ncl, resolution ….
    - Derived "invariant" variables:
        - e.g RMS of current map/phi averaged map/scaled maps
        - distortion map/phi averaged map
        - local discontinuities in time and space
        - "physics" acceptable performance

Explain/find hardware origin of anomalies (currently hierarchy of alarms)

Training data:

- current maps, and distortion maps (at GSI) can be exported to alien
- time bin based QA currently available only for few run

DCAr alarms decomposition

dEdx alarms decomposition