Universidade de São Paulo
Instituto de Física

Desenvolvimento de algoritmo de clusterização para calorímetro frontal do experimento ALICE no LHC

Danilo Anacleto Arruda da Silva

Orientador: Prof. Dr. Nelson Carlin Filho

Dissertação de mestrado apresentada ao Instituto de Física para a obtenção do título de Mestre em Ciências

Banca Examinadora:

Prof. Dr. Nelson Carlin Filho (Instituto de Física)
Prof. Dr. Prof. Dr. Marco Aurélio Lisboa Leite (IFUSP)
Prof. Dr. Prof. Dr. Joel Mesa Hormaza (UNESP)

São Paulo
2014

-
-
-

FICHA CATALOGRÁFICA
Preparada pelo Serviço de Biblioteca e Informação do Instituto de Física da Universidade de São Paulo

Silva, Danilo Anacleto Arruda da

Desenvolvimento de algoritmo de clusterização para calorímetro frontal do experimento ALICE no LHC. São Paulo, 2014.

Dissertação (Mestrado) – Universidade de São Paulo. Instituto de Física. Departamento de Física Nuclear

Orientador: Prof. Dr. Nelson Carlin Filho

Área de Concentração: Física Nuclear de Altas Energias.

Unitermos: 1.Física de alta energia; 2. Métodos de clusterização; 3. Calorimetria; 4. Color glass condensate.

USP/IF/SBI-084/2014

"O meu coração ferve com palavras boas; falo do que tenho feito no tocante ao Rei; a minha língua é a pena de um destro escritor."

Salmos, 45:1

Agradecimentos

Agradeço a Deus por todas as coisas que Ele tem me dado e me ajudado.
Ao meu Pai por ser o meu Herói e a pessoa em quem me espelho.
A minha Mãe por sempre ter desejado o meu bem durante este período.
A minha Irmã por não ter deixado meu lado "nerd" morrer.
A Mulher da minha vida, Daniele, por ter me ajudado, incentivado e segurado as minhas mãos nos meus momentos mais difíceis.
Ao meu orientador Prof. Dr. Nelson Carlin pela paciência.
Ao Prof. Dr. Marcelo Munhoz por ter me ajudado em várias partes deste trabalho.
Ao Caio Eduardo e Juliana Raw pelas boas conversas, noites jogando Magic e pelas "jogatinas" aos sábados a tarde.
Ao Caio Prado por ter me salvado várias vezes dos problemas em meu códigos.
Ao Renato Negrão e a Camila pela ajuda em alguns momentos deste mestrado.
Ao Viktor Jahnke por ter me ajudado em várias disciplinas da pós-graduação.
Ao pessoal do Judô do CEPEUSP (Sensei Pascoal Tambucci, Takamoto, Miroslav, Capelle e Vítor) pelos excelentes treinos.
E a todos os meus amigos que não tiveram seus nomes aqui inscritos.

Resumo

O Grande Colisor de Hádrons (*Large Hadron Collider* - LHC) é um acelerador de prótons e íons pesados localizado no CERN (*Conseil Européen pour la Recherche Nucléaire*). Em um de seus experimentos, ALICE (*A Large Ion Collider Experiment*), está sendo projetado um detector dedicado a explorar os aspectos únicos de colisões núcleo-núcleo. A principal finalidade do ALICE é estudar a formação de um novo estado da matéria, o plasma de quarks e glúon. Para isto devem-se ter medidas precisas de hádrons, elétrons, múons e γ produzidos em colisões chumbo-chumbo. Assim está sendo proposto um calorímetro frontal (*Foward Calorimeter* - FoCal) como um *upgrade* para o ALICE. A função deste calorímetro é o estudo das funções de distribuição de pártons (*Partons distribuction Functions* - PDF) no regime de pequenos valores do x de Bjorken. Nesta região é esperado que estas PDFs tenham um comportamento não linear devido ao processo de saturação de glúons. Para o estudo desta região é necessária a medida de fótons diretos produzidos na colisão. Estes, por sua vez, ficam mascarados pelo fundo de fótons provenientes do decaimento de π^0, o que leva a uma necessidade de suas identificações. Com isto surge a oportunidade para a utilização do método de clusterização que é uma ferramenta de mineração de dados. Este trabalho contribuiu para o desenvolvimento inicial de um algoritmo de clusterização para o calorímetro FoCal.

Palavras-chave: *Color Glass Condensate*, Calorimetria, Métodos de Clusterização.

Abstract

The Large Hadron Collider (LHC) is a CERN's a accelerator that collides protons and heavy ions. One of its experiments, ALICE, is building a new detector to explore new aspects of heavy ions collisions. The Alice's main goal is to study the formation of quark-gluon plasma (QGP). To do that it's necessary to get accurate data on hadrons, electrons, muons and gammas created in lead-lead collision. So, to accomplish that a new calorimeter is in study to scan the foward region of experiment, the Foward Calorimeter (FoCal). It's proposed to study Parton Distribution Functions (PDFs) in a regime of very small Bjorken-x, where it is expected that the PDFs evolve non-linearly due to the high gluon densities, a phenomena referred to as gluon saturation. But to do that it's required to measure the direct gammas created on collision. These fotons are blended on by fotons descendant of π^0. So there's a need to separate it from the direct gammas. One way to solve this problem is to use clustering methods (a type of mining data algorithm). This work helped on early stages of development that clustering algorithm.

Key-words: *Color Glass Condensate*, Calorimetry, Clustering Method.

Sumário

Lista de Figuras

Capítulo 1

Introdução

O Grande Colisor de Hádrons (*Large Hadron Collider - LHC*) é um colisor de prótons e íons pesados localizado na fronteira entre a França e a Suíça que está em operação desde 2010. Ele é composto por quatro experimentos, a saber: ATLAS (*A Toroidal LHC Apparatus*), CMS (*Compact Muon Solenoid*, LHCb (*Large Hadron Collider beauty*) e ALICE (*A Large Ion Collider Experiment*). Os dois primeiros experimentos se concentram na busca pelo bóson de Higss, dimensões extras e partículas que podem ser parte da matéria escura. O experimento LHCb investiga a pequena diferença entre matéria e antimatéria através do estudo da partícula chamada *quark beauty*. O experimento ALICE estuda a fase da matéria chamada de Plasma de Quarks e Glúons (*Quark-Gluon Plasma* - QGP), sendo que dentro deste é onde se concentra este trabalho

Apesar de estar há quatro anos em operação, já estão em curso ações para seu *upgrade*. Para o experimento ALICE, um dos *upgrades* é o desenvolvimento de um calorímetro frontal (*Foward Calorimeter* - FoCal). Este por sua vez irá contribuir para o estudo das funções de distribuições de pártons (*Parton Distribuction Function* - PDF), para núcleos, na região de pequenos-x. Nesta região ocorre a saturação de glúons que é descrito pela teoria do *Color Glass Condensate* (CGC).

As partículas medidas na regiões frontais sofrem um *boost* longitudinal, no caso de mésons neutros (píons), as partículas de seus decaimentos terão ângulos pequenos devido a este *boost*, o que leva a uma união de seus chuveiros eletromagnéticos no calorímetro. A separação destes chuveiros é importante pois o maior interesse está em se medir as partículas provenientes das interações que ocorrem entre os quarks e glúons e não as provenientes de decaimento. Esta separação pode ser realizada utilizando detectores de pixels e/ou algoritmos de mineração de dados.

Os algoritmos de mineração que são utilizados para classificação e separação de dados são os de aprendizado de máquina. Estes podem ser separados em duas categorias: aprendizado supervisionado e não supervisionado. Para o primeiro grupo temos como um exemplo as redes neurais e para o segundo os algoritmos de clusterização (particiona-

mento, fuzzy, etc). No aprendizado supervisionado tem-se uma clara medida do sucesso ou fracasso de sua classificação, pois existe uma maneira se comparar seu resultado obtido com um tido como verdadeiro. Para o caso não supervisionado isto não é possível, pois não há um resultado tido como verdadeiro, então a classificação é realizada utilizando somente as informações contidas no seu conjunto de dados, e sua validade é definida pelo investigador.

1.1 Objetivos

Observando o exposto acima, o objetivo deste trabalho é contribuir para o desenvolvimento do FoCal. Isto será realizado através do estudo de algoritmos de aprendizado não supervisionado, especificamente a clusterização.

1.2 Organização do Trabalho

Este trabalho está organizado da seguinte maneira: no capítulo 1 discorre a presente introdução ao trabalho. No capítulo 2 é apresentada uma revisão do acelerador LHC e do experimento ALICE. No capítulo 3 é apresentada uma revisão sobre o funcionamento de um calorímetro. No capítulo 4 é explicado o que é o FoCal, qual a motivação para seu desenvolvimento e o que ele vai medir. No capítulo 5 serão apresentados os aspectos teóricos da clusterização e explicações de seus principais algoritmos. Nesta seção também é descrito o algoritmo utilizado neste trabalho. No capítulo 6 são apresentados e discutidos os resultados obtidos para três configurações de geometria e por último no capítulo 7 é apresentada a conclusão deste trabalho.

Capítulo 2

O Acelerador LHC

Um dos principais temas a serem abordados no Large Hadron Collider (LHC) no CERN é a conexão entre transições de fase envolvendo campos quânticos elementares, simetrias fundamentais da natureza e a origem das massas. O programa experimental do LHC vai explorar esses aspectos por meio de abordagens experimentais complementares de pelo menos quatro grandes experimentos. Os experimentos ATLAS[1] e CMS[2] se concentram no Bóson de Higgs. O experimento LHCb[3] vai estudar processos de violação de simetria. O experimento ALICE vai investigar o papel da simetria quiral na geração da massa em hádrons utilizando colisões entre íons-pesados para atingir elevadas densidades em largos volumes e elevadas escalas temporais. Podem ser estudados fenômenos de equilíbrio e não equilíbrio num largo intervalo de densidades. Espera-se também reunir informações adicionais sobre a estrutura do diagrama de fase da QCD. O primeiro trabalho publicado com dados do LHC foi do experimento ALICE, apresentando estudo da densidade de pseudo-rapidez para partículas carregadas primárias em colisões p-p a $\sqrt{s} = 900GeV$[4]. Os resultados se mostraram consistentes com resultados anteriores de colisões entre p e antipróton obtidos no CERN na mesma energia.

O LHC (Large Hadron Collider)[5] está instalado no CERN em um grande túnel subterrâneo com 27 km de extensão. O processo de aceleração começa no acelerador LINAC 2 para o caso de prótons, e no LINAC 3 para íons de chumbo. Os prótons acelerados no LINAC 2 são injetados no Proton Synchrotron Booster, com energia de 50 MeV, de onde saem com energia de 1.4 GeV. Em seguida são injetados no Proton Synchrotron (PS), onde adquirem uma energia de 26 GeV. O próximo estágio corresponde à injeção no Super Proton Synchroton (SPS), onde os prótons adquirem energia de 450 GeV, sendo então injetados no LHC, no qual cada feixe viajando em sentidos opostos pode atingir até 7 TeV.

O feixe de chumbo segue caminhos diferentes no início do processo. O LINAC 3 cria feixes com 4.2 $MeV/ncleon$, os quais são injetados Low Energy Injection Ring (LEIR), que produz íons com 72 $MeV/ncleon$. Estes passam então pelo PS e SPS, sendo então

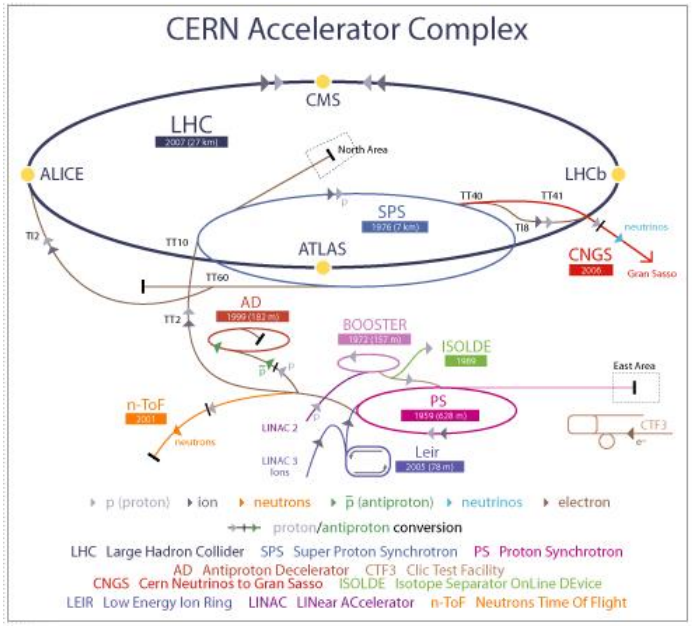


Figura 2.1: Esquema do complexo de aceleradores do CERN.

injetados no LHC, podendo chegar até a 2.8 $TeV/ncleon$.

O LHC é formado por 1232 dipolos magnéticos supercondutores que podem operar com campos magnéticos de até 9 T. Também compõem o sistema, mais de 500 quadrupolos magnéticos e mais de 4000 ímãs para correção de trajetória de feixe ao longo do anel. O projeto do acelerador prevê que pode-se atingir uma luminosidade de 1034 $s^{-1}cm^{-1}$

Atualmente são 4 os experimentos em operação no LHC: ALICE, ATLAS, CMS e LHCb. Esses experimentos estão localizados em pontos definidos ao longo do anel. Na Figura 2.1 é mostrado um esquema do complexo de aceleradores do CERN, usados para a obtenção dos feixes de prótons e chumbo. A localização dos 4 experimentos é também mostrada.

No final de 2009, após uma parada devido a acidente criogênico, as primeiras colisões forma registradas a uma energia de $\sqrt{s} = 900GeV$ e em seguida, $2,4TeV$. Em 30 de março de 2010, foram realizadas as primeiras colisões com $\sqrt{s} = 7TeV$.

2.1 Experimento ALICE

O experimento ALICE foi idealizado com o objetivo de estudar a criação de um plasma de quarks e glúons[6][7]. Com as elevadas energias atingidas pelo LHC, podemos obter valores elevados de densidade, tamanho e tempo de vida da matéria de quarks excitada, permitindo investigações detalhadas desse estado da matéria. Nesse caso, a temperatura vai exceder significativamente o valor crítico previsto para a transição. O experimento

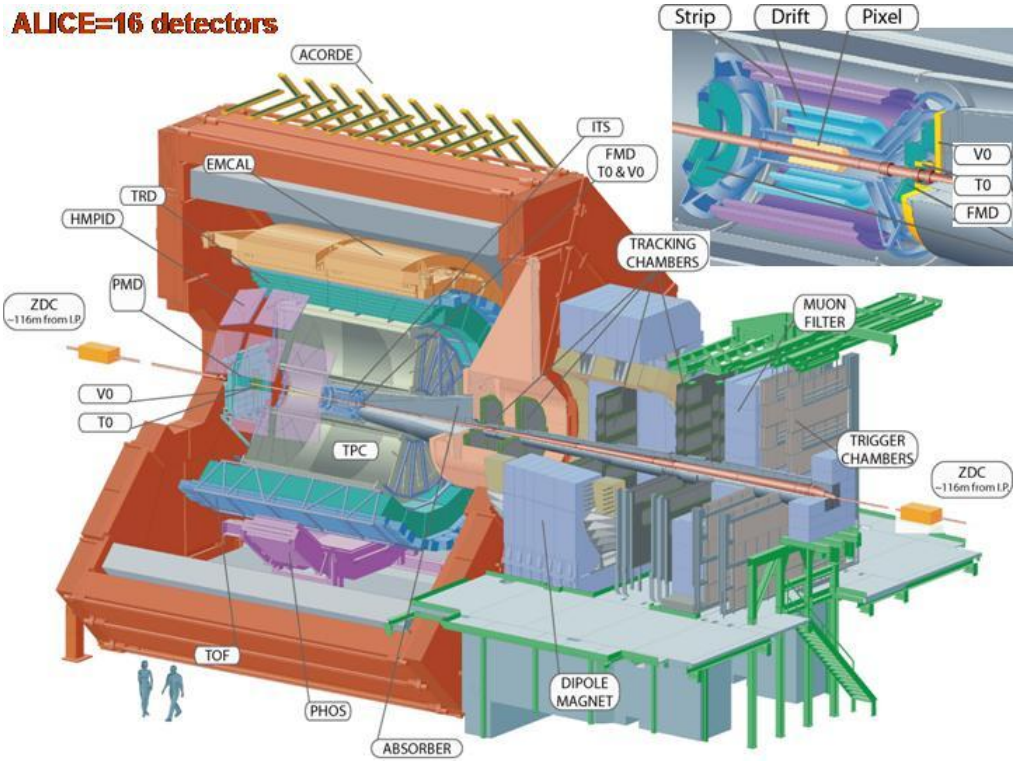


Figura 2.2: Esquema do experimento ALICE.

ALICE é composto por 16 detectores. Na Figura 2.2 é mostrado um esquema do experimento com os principais elementos. Uma descrição detalhada do experimento é dada em [8].

O experimento possui um sistema de tracking cobrindo o intervalo de momento transversal de $100MeV/c$ a $100GeV/c$ e capacidade de identificação para píons, káons, prótons, elétrons, múons e fótons. Os múons são detectados na direção frontal por meio de um conjunto de câmaras para tracking no interior de um dipolo magnético. A parte central é inserida no interior de um solenóide com campo magnético de 0.5 T. Os sistemas de tracking central e identificação de partículas cobrem um intervalo de pseudo-rapidez de $-0.9 < \eta < 0.9$. A medida das trajetórias é realizada por meio de um conjunto de 6 barris com detectores de Si, que compõem o Inner tracking System (ITS), e uma Time Projection Chamber (TPC), com um volume de 88 m^3. Com essa configuração, pode-se obter perdas de energia e efetuar a identificação de partículas. Um Transition Radiation Detector (TRD) e um sistema de tempo de vôo (TOF) fornecem identificação em momentos intermediários. A identificação de partículas com momentos mais elevados pode ser realizada com o auxílio de um detector do tipo Ring Imaging Cherenkov (HMPID), cobrindo 15% da área central. Com isso, pode-se separar píons de káons com momentos até $3GeV/c$ e káons de prótons até $5GeV/c$. Um calorímetro eletromagnético (EMCAL) vai cobrir o anel central. Elétrons e fótons são medidos no detector PHOS que consiste de um

calorímetro de alta resolução. Um calorímetro na região frontal (ZDC) ajuda a complementar informações sobre a multiplicidade das partículas.Um detector proporcional para medida da multiplicidade de fótons(PMD) está instalado na região frontal em dos lados do conjunto. Em cada lado do ponto de interação há um sistema de cintiladores (V0), utilizado como o trigger principal para interação. Na parte superior do ímã situa-se um detector para raios cósmicos (ACORDE), o qual identifica a chegada de múons cósmicos. Com essa configuração do projeto, pretende-se obter informações sobre a densidade de energia atingida. As medidas de fluxo elíptico fornecem informações sobre termalização e equação de estado do sistema na fase de elevada densidade. A medida das razões entre a produção de partículas no estado final está relacionada ao equilíbrio químico, o que pode permitir traçar a trajetória do sistema no diagrama de fase. A evolução espaço-temporal do sistema pode ser investigada por meio da utilização de métodos de interferometria, incluindo o estudo de ressonâncias. O estudo das abundâncias de J/ψ e Upsilon podem fornecer informações sobre deconfinamento, assim como o estudo da produção de jatos pode propiciar a investigação das propriedades de transporte de pártons espalhados no meio, as quais podem ser muito modificadas se um plasma de quarks e glúons é formado.

Capítulo 3

Calorimetria

Métodos calorimétricos em física nuclear ou de partículas implicam absorção total da energia da partícula num volume do material, seguida pela medida da energia depositada. Por exemplo, múons de alta energia perdem energia principalmente,por ionização. Por outro lado, fótons de altas energias, elétrons e hádrons, interagem com o meio produzindo partículas secundárias, que levarão ao desenvolvimento de um chuveiro (cascata) eletromagnético. Desta maneira a energia da partícula é depositada no material de uma maneira muito mais eficiente. Devido a isso, os calorímetros são usados largamente como detectores em física de altas energias para detectar chuveiros hadrônicos ou eletromagnéticos. Tais detectores são chamados de calorímetros eletromagnéticos ou hadrônicos[9].

3.1 Chuveiro Eletromagnético

3.1.1 Energia Perdida por Partículas Carregadas

O mecanismo melhor conhecido de perda de energia é a interação eletromagnética sofrida por partículas carregadas. Essas interações são[10]:

Ionização Liberação de um elétron atômico do campo coulombiano produzido pelo núcleo.

Excitação Partículas carregadas podem excitar átomos/moléculas sem ionizá-las. Ao retornar do estado excitado, este átomo/ molécula emite luz (cintilação), a qual é medida.

Radiação de Cherenkov Partículas atravessando um meio com velocidade superior a da luz (neste meio) perdem energia por radiação de Cherenkov.

Em altas energias Partículas pedem energia por raios δ, *bremsstrahlung* ou indução de reações nucleares.

Em energias superiores a 100 MeV, elétrons e pósitrons perdem energia principalmente por *bremsstrahlung*. Este processo ocorre quando o elétron (ou pósitron) interage com o campo coulombiano do núcleo, resultando na emissão de fóton que, geralmente, carrega uma pequena fração da energia do elétron (pósitron). O espectro de energia destes fótons cai com $1/E$[10][9].

Neste processo o elétron (ou pósitron) sofre um pequeno desvio em sua trajetória, que é chamado de espalhamento múltiplo ou espalhamento Coulomb. Este depende da energia do fóton emitido e do Z do material (força do campo de Coulomb)[10].

Estes processos citados acima ocorrem para todas as partículas carregadas atravessando um meio, mas para ocorrerem, as energias destas partículas devem estar acima de uma energia mínima definida como energia crítica, E_c. Ela é definida como aquela em que a energia perdida por ionização a cada comprimento de radiação (X_0)[1] se iguala a energia do elétron[10]

$$(\Delta E)_{ion} = \left[\frac{dE}{dx}\right]_{ion} X_0 = E \tag{3.1}$$

Uma outra definição toma como a energia em que a média de perda de energia por processos radiativos se iguala àquelas por ionização. Assim, a primeira definição é igual a esta se a perda por *bremsstrahlung* é dada por [10][11]

$$\left[\frac{dE}{dx}\right]_{brems} = \frac{E}{X_0} \tag{3.2}$$

A qual é válida para altas energia em que a perda por ionização é desprezível, sendo que esta é apenas uma aproximação em regimes próximos a E_c.

3.1.2 Interação por Fótons

Os fótons interagem com a matéria através de cinco processos[10]:

Efeito Fotoelétrico Ele é o processo que tem maior chance de ocorrer a baixas energias. Neste caso o átomo absorve um γ e emite um elétron. O átomo fica então excitado e decai emitindo elétron *auger* ou raios-X. A seção de choque deste processo depende do número de elétrons disponíveis, e assim do Z do material. Esta também varia com a energia do γ como E^{-3}, e conforme o aumento da energia perde importância rapidamente.

[1] veja seção 3.2

Espalhamento Coerente (Rayleigh) É um processo importante a baixas energias. O γ é defletido por um elétron atômico, mas não perde energia. Este mecanismo afeta a distribuição espacial de energia depositada, mas não contribui para a deposição em si.

Espalhamento Incoerente (Compton) O fóton é espalhado por um elétron atômico, com transferência de energia e momento suficientes para colocá-lo num estado não ligado. Este processo é mais provável de ocorrer em materiais de alto Z, e pelo menos metade da energia é depositada por este mecanismo.

Produção de Pares Fótons com $E > 2m_e c^2$, na presença de um campo produzido por uma partícula carregada, pode criar um par elétron-pósitron. 99% das vezes que ocorre este processo, o fóton está na presença de campos magnéticos nucleares. Este processo depende fortemente da energia do fóton incidente e da densidade de elétrons ($\sim Z$) do meio.

Reações Fotonucleares Em energias de 5 a 20 MeV um pequeno papel é desempenhado por reações fotonucleares, tipo γn, γp ou fissões induzidas por fótons.

3.1.3 Exemplo de um Chuveiro Eletromagnético

A característica mais importante de um chuveiro pode ser compreendida em um modelo bem simples. Seja E_0 a energia de um fóton incidente num material (figura 3.1).

Após um comprimento de radiação o fóton produz um par e^+e^-. Elétrons e pósitrons emitem um fóton de *bremsstrahlung* depois de outro comprimento de radiação, o qual irá novamente produzir um outro par. Assumindo-se que a energia é simetricamente dividida entre as partículas em cada etapa da multiplicação, o número de partículas no chuveiro (soma de elétrons, pósitrons e fótons) na profundidade t é dado por

$$N(t) = 2^t \tag{3.3}$$

em que a energia das partículas individuais na geração t é

$$E(t) = E_0 2^{-t} \tag{3.4}$$

A multiplicação do chuveiro continua até enquanto $E_0/N > E_c$. Quando a energia da partícula fica abaixo do valor crítico E_c processos de absorção como ionização, para elétrons, e Compton e fotoelétricos para fótons começam a dominar. A posição onde o

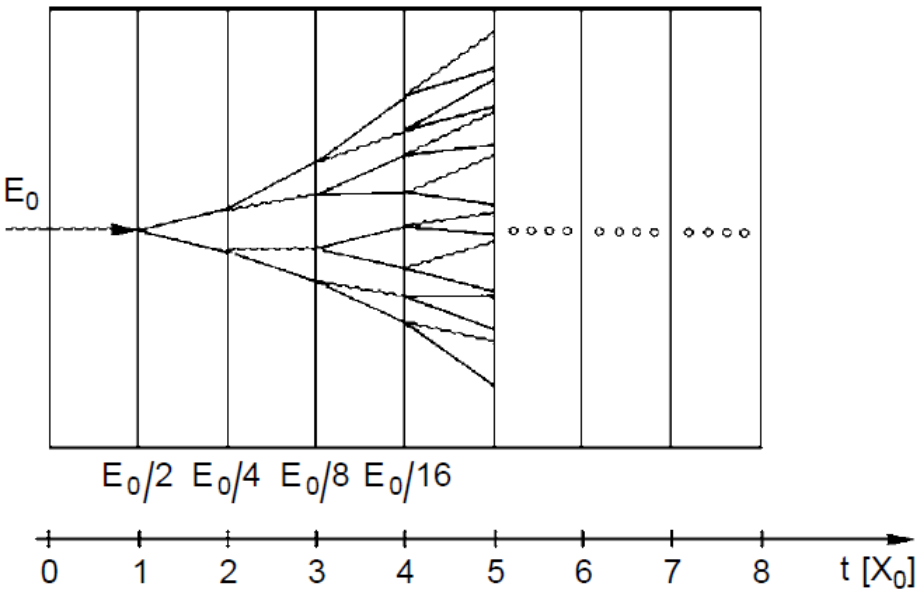


Figura 3.1: Esquema de um chuveiro eletromagnético. Retirado de [9].

chuveiro é máximo (*shower maximum*) é alcançado quando

$$E_c = E_0 2^{-t_{max}} \tag{3.5}$$

isto leva a

$$t_{max} = \frac{\ln(E_0/E_c)}{\ln 2} \propto \ln(E_0/E_c) \tag{3.6}$$

Após o *shower maximum* os elétrons irão parar depois de 1 X_0. Os fótons de mesma energia irão penetrar muito mais no detector[9]. Este modelo simples descreve corretamente as características mais importantes de uma cascata eletromagnética. Com isso tem-se que[9]:

- Para absorver a maior parte da energia de um fóton incidente a espessura total de um calorímetro deve ter pelo menos 10-15X_0.
- A posição do *shower maximum* varia muito pouco com a energia. Portanto a epessura do calorímetro deveria aumentar logaritimamente com a energia (para múons será proporcionalmente).
- A fuga de energia (*Energy leakage*) ocorre principalmente por fótons de baixa energia que escapam do calorímetro pelas laterais (*lateral leakage*) ou pela parte de trás (*longitudinal leakage*).

Na realidade o desenvolvimento do chuveiro é muito mais complexo. Isto está mostrado na figura 3.2.

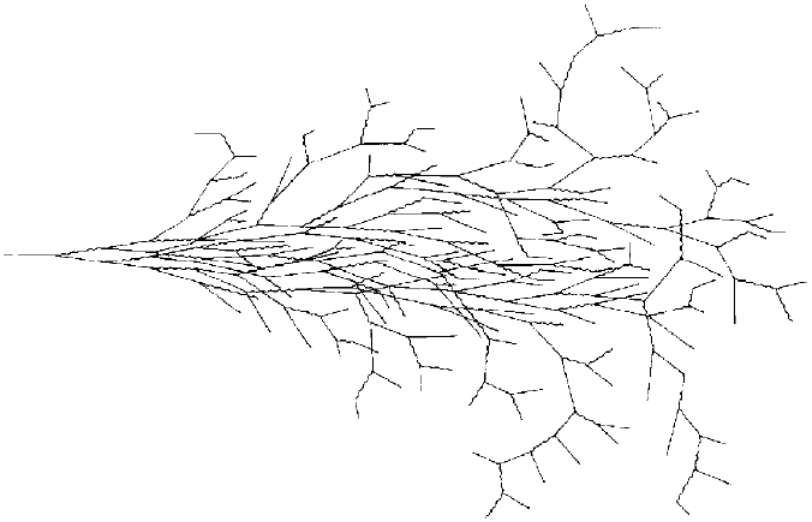


Figura 3.2: Representação esquemática de um chuveiro eletromagnético. Retirado de [9].

3.2 Variáveis de Escala

Comprimento de Radiação (X_0) Definido como a distância na qual um elétron ou pósitron de alta energia ($>> 1\ GeV$) perde, em média, 63% ($1 - e^{-1}$) de sua energia devido a *bremsstrahlung*. Ao expressar as dimensões do absorvedor em unidades de X_0, efeitos que dependem do material são em primeira aproximação eliminados. Para cálculos aproximados (precisão de 3%) utiliza-se a equação abaixo[10],

$$X_0 = \frac{716,4A}{Z(Z+1)\ln(287/\sqrt{Z})}\ g\ cm^{-1} \tag{3.7}$$

Raio de Molière Esta quantidade não tem um significado igual em precisão ao do comprimento de radiação; O raio de Molière é frequentemente usado para descrever o desenvolvimento transversal do chuveiro de uma maneira aproximadamente independente do material. Ele é definido em função da energia crítica e de X_0, como segue:

$$\rho_M = E_s \frac{X_0}{E_c} \tag{3.8}$$

em que a escala de energia E_s, definida como $m_e c^2 \sqrt{4\pi/\alpha}$, é igual a 21,2 MeV. Em média 90% do chuveiro é depositado num cilindro de raio ρ_M em torno do eixo do chuveiro[10].

3.3 Fuga de Energia

O calorímetro é instrumento utilizado para se medir as propriedades das partículas através de sua absorção total. Na prática, este total significa 99.9%, ou 99%, ou menos. O que indica que o chuveiro não é completamente absorvido (contido) dentro do volume do detector. Exitem dois aspectos deste problema. Primeiro, quando não se contém completamente o chuveiro, isto leva a flutuações por evento relacionadas à medida da energia, o que afeta a qualidade da informação obtida. Segundo, não conter o chuveiro significa que as partículas escapam do calorímetro, e estas podem causar sinais em outros detectores[10].

A fuga é um efeito que depende da energia. A fração desta que é carregada pela partícula do chuveiro que não é depositada, depende (além da energia)do tipo de partícula. Dado um calorímetro, elétrons com uma determinada energia são melhores absorvidos que prótons, que por sua vez são melhores contidos que píons[10]. Sendo assim, abaixo estão os três tipos de fuga que podem ocorrer,

Fuga Longitudinal (*Longitudinal Leakage*) Partículas do chuveiro escapam da detecção por saírem pela a parte traseira do calorímetro. Considerações sobre este tipo de fuga, frequentemente, dita o *design* do calorímetro pois seu comprimento impacta diretamente no custo.

Fuga Lateral (*Lateral Leakage*) Apesar da preocupação com a profundidade do calorímetro, na prática são os efeitos da fuga lateral que dominam. Para determinar a energia da partícula, tende-se a limitar a área ao redor do eixo do chuveiro na qual o sinal do calorímetro é integrado, levando a perdas laterais que são uma ordem de grandeza superior às longitudinais.

Albedo É a fuga através da superfície frontal. Este tem um papel significante em energias muito baixas. As partículas que escapam desta maneira têm energias muito baixas e são produzidas em espalhamentos nos quais elas têm ângulos grandes. Exemplos de tais processos, são o espalhamento Compton e o efeito fotoelétrico, em caso de chuveiros eletromagnéticos e espalhamento elástico com nêutrons em caso de chuveiros hadrônicos. Dos três tipos, este é o único que é inevitável não importando o tamanho do calorímetro.

3.4 Calorímetro Homogêneo

Calorímetros homogêneos são construídos de um material que combina as propriedades de um absorvedor e detector. Isto significa praticamente que todo o volume do calorí-

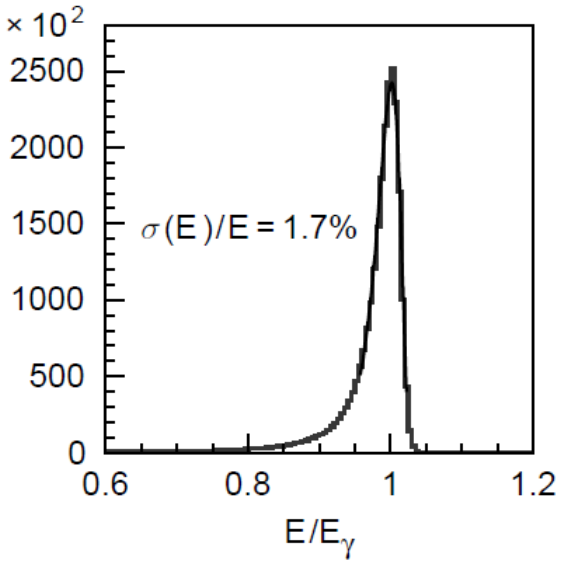


Figura 3.3: Espectro de energia medido com calorímetro para fótons de 4-7 GeV. Retirado de [9].

metro é sensível a energia depositada. Estes calorímetros são baseados na medida de luz da cintilação emitida por esses materiais (cristais cintiladores, gases nobres liquefeitos), ionização (gases nobres líquidos) e radiação de Cherenkov (*lead glass* ou cristais pesados transparentes)[11][9].

O principais parâmetros para um calorímetro eletromagnético são a resolução em energia e posição para fótons e elétrons. A resolução em energia (σ_E/E) é determinada por fatores físicos (flutuação na fuga da energia ou estatística do efeito fotoelétrico) e técnicos (não uniformidade dos cristais)[11][9].

Um típico espectro de energia medido num calorímetro é mostrado na figura 3.3. Para um sistema de alta resolução esta curva é assimétrica, com uma grande "cauda" em baixas energias, e a resolução em energia é dada como

$$\sigma_E = \frac{FWHM}{2.35} \tag{3.9}$$

Este espectro tem esta forma devido a diferentes interações que podem ocorrer no interior do detector. Como um exemplo, considere um feixe monoenergético de partículas carregadas (p.e. elétrons) incidindo num detector espesso o suficiente para parar as partículas. Assumindo que todos os elétrons irão perder energia por colisões atômicas os espectro será uma gaussiana. Na realidade, alguns elétrons irão ser espalhados para fora do detector antes de depositarem completamente sua energia. Isto produz uma cauda de baixa energia. Outros elétrons irão produzir fótons de *bremsstrahlung* que po-

dem escapar do detector. Isto, novamente produz eventos com energia menor que o pico da gaussiana[12]. Esta distribuição assimétrica está mostrada na figura 3.3 e pode ser aproximada por uma distribuição log-Normal. O motivo para a escolha desta função é o seguinte: imagine uma variável aleatória X que é a soma de várias variáveis aleatórias f_i, portanto

$$X = \sum_{i=1}^{n} f_i \tag{3.10}$$

pelo teorema central do limite, quando $n \to \infty$, a f.d.p. de X tende a ter uma distribuição normal. Observando o calorímetro, temos que a energia depositada E_d é um produto de várias frações f_j (estas dependem de como as partículas caregadas ou não perdem energia no calorímetro) da energia incial E_i, portanto

$$E_D = \prod_{j=1}^{n} f_j E_i \tag{3.11}$$

Agora tomando o logaritmo de ambos os lados desta equação, têm-se

$$\ln(E_D) = \sum_{j=0}^{n} \ln(f_j) + \ln(E_i) \tag{3.12}$$

Com isto quando $n \to \infty$ a variável $\ln(E_D)$ tem distribuição normal.

A f.d.p. log-normal é dada por:

$$f(x; \mu, \sigma) = \frac{A}{x\sigma} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}} \tag{3.13}$$

os parâmetros de ajuste são A, μ e σ.

3.5 Calorímetros Segmentados

Há uma maneira mais simples e econômica de se medir a energia do fóton se a resolução em energia não for crucial. Voltemos ao modelo simplificado do chuveiro e coloquemos uma fina camada de um contador na região onde se tem o maior número de partículas produzidas (*shower maximum*). Neste modelo simples o número de partículas atravessando o contador (veja fórmulas 3.5 e 3.6) é 2/3 do $N_{max} = E_\gamma/E_c$, porque N_{max} é dividido igualmente entre fótons, elétrons e pósitrons. A amplitude do sinal do contador é proporcional ao número de partículas carregadas.

Para tomar vantagem da ideia discutida, normalmente se desenha um calorímetro como um conjunto de contadores finos separados por absorvedores. Estes tipos de calorímetros são chamados de calorímetros segmentados (*sampling calorimeters*) já que esses

medem somente uma parte (amostra) da energia depositada. Em adição à fuga de energia (*energy leakage*), a resolução em energia destes calorímetros é afetada por flutuações da amostragem [9].

Como o principal objetivo dos calorímetros é a medida da energia, a característica física mais importante é a resolução em energia. Considerando um calorímetro segmentado, o chuveiro se desenvolve nas camadas de alto Z (maior parte da energia é perdida) e sua energia é medida nas camadas de material de baixo Z (p.e. cintiladores, plásticos ou argônio líquido). A parte da energia medida deve ser proporcional à energia absorvida. O modelo básico para um calorímetro segmentado é onde o chuveiro se desenvolve no material inerte de alto Z e a energia é medida na parte ativa de baixo Z[9][11].

Um problema que ocorre neste calorímetro é que muitos elétrons de baixa energia (originados por fótons ao interagir com a matéria) não são medidos, pois estes são produzidos na camada absorvedora e não chegam ao material ativo. Como estes são em maior quantidade num chuveiro a medida da energia não é computada adequadamente[10].

Capítulo 4

FoCal - *Foward Calorimeter*

4.1 O que é?

O FoCal é um calorímetro eletromagnético que está sendo proposto como um *upgrade* para o detector ALICE [13]. Este, por sua vez, é um projeto envolvendo vários países, como por exemplo: Brasil, Holanda, Japão, índia, EUA e República Tcheca. Alguns deste estão envolvidos na construção de protótipos que foram testados em Novembro de 2011 (p.e. Holanda e Japão) e outros com a simulação (p.e. Brasil, Holanda, Japão e República Tcheca).

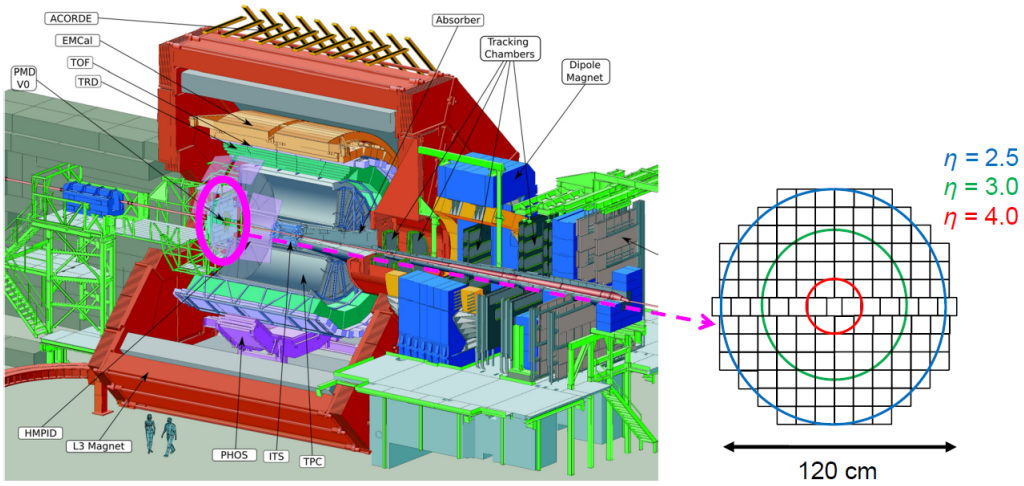
Em 2011 o FoCal estava sendo projetado para ser um calorímetro com 177 torres que operam independentemente. Ele fica a 3,5m do ponto de colisão, no lugar do PMD, e abrange rapidez de 2,5 a 4,5 com total aceitação azimutal. Cada torre é altamente segmentada, entre 24 e 30 camadas (*layers*), e composta por uma combinação de folhas de tungstênio mais um sensor de silício. Para uma representação pictórica deste calorímetro, veja a figura 4.1.

Neste período estava-se estudando dois tipos de tecnologia de detecção para o calorímetro:

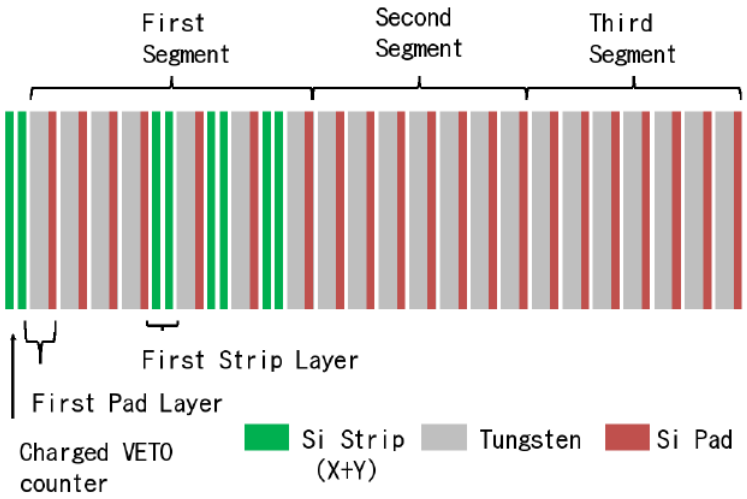
1. Sensor discreto - são necessários dois tipos de sensores: *pad* e *strip*[9]. A maioria das camadas são equipadas com sensores do tipo *pad*, ao passo que somente 5 camadas (incluindo a primeira) são de um sensor *strip* de dupla camada, ou dois sensores *strip* juntos numa camada.

2. Sensores de pixel integrado (MIMOSA26) - este tipo de sensor esta sendo desenvolvido dentro do projeto europeu EDUNET. Sua matriz irá cobrir uma área de 224 mm^2, e está organizada em 576 linha por 1152 colunas. Estima-se que o tempo de aquisição de cada evento seja de 100 μs.

Nesta época a proposta do calorímetro não se encontrava madura, devido ao estágio



(a)



(b)

Figura 4.1: (a) Esquema do calorímetro FoCal (geometrias 1 e 2) visto de frente e sua localização no detector ALICE. (b) Esquema de uma única torre.

inicial em que se encontrava o desenvolvimento, era possível realizar algumas variações na geometria apresentada na figura 4.1. A variações nesta geometria realizadas neste trabalho foram chamadas de Geometrias 1 e 2. Suas descrições e resultados obtidos com elas estão apresentados nas seções, respectivamente, 6.1 e 6.2.

Com o desenvolvimento das simulações e análises executadas por outros membros da colaboração, chegou-se a propor uma nova configuração para o FoCal (FoCal-E) e iniciaram-se discussões para a implementação de um calorímetro frontal hadrônico (FoCal-H). Neste trabalho só foi estudado o comportamento do algoritmo de clusterização para o calorímetro eletromagnético (FoCal-E). Segue uma descrição de ambos os calorímetros.

FoCal-E O calorímetro FoCal eletromagnético (FoCal-E) será segmentado utilizando como meio absorvedor o tungstênio (W) e como meio para leitura da energia o silício (Si). Para o silício serão usadas duas tecnologias:

- *Layers* de baixa granularidade (*Low Granularity Layer* - LGL) com células de tamanho $\approx 1cm^2 \approx R_M^2$, o qual são combinados longitudinalmente em segmentos e
- *Layers* de alta granularidade (*High Granularity Layer* - HGL) com tamanho de $\approx 1mm^2$.

Uma visão esquemática do detector está apresentada na figura 4.2. Todos os *layers* consistem de folhas de W de $\approx 1X_0$ seguida por uma camada ativa de silício. Esta figura mostra uma estrutura com 3 segmentos LGL, cada um somado e lido independentemente, com 2 camadas de HGL colocadas nas camadas 5 e 10. O local para o HGL ainda está em estudo, mas este provavelmente é o design ótimo. A provável localização do calorímetro será no lugar do PMD.

A tecnologia de sensor de silício a ser utilizada é o *monolithic active pixel sensors* (MAPS)[1]. Como estes são baseados na tecnologia CMOS eles serão relativamente baratos. A tecnologia mais avançada pode prover sinais binários em pixels de tamanho de $20 \times 20\mu m^2$. Neste design estes pixels serão colocado juntos logicamente em macro-pixels, ou mini-pads, de $1 \times 1mm^2$. A contagem destes mini-pads serão como sinais efetivos do HGLs. Os resultados obtidos para esta geometria, denominada de Geometria 3 neste trabalho, estão apresentados na seção 6.3.

FoCal-H Calorímetro Hadrônico Opcionalmente ao FoCal-E será complementado por calorímetro hadrônico. Idealmente ele irá cobrir a mesma faixa de pseudo-rapidez do FoCal-E e deverá ficar o mais próximo possível do deste. Sua localização será a 10

[1]Para mais detalhes sobre o desenvolvimento e características de sensores de pixels, vide apêndice A

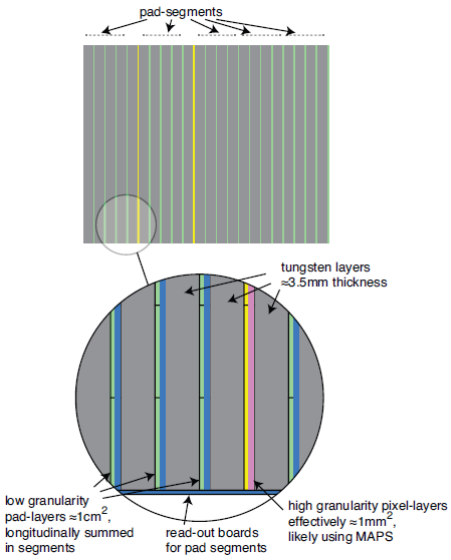


Figura 4.2: Visão esquemática da estrutura longitudinal do FoCal-E (geometria 3).

m do ponto de interação, devido ao suporte e região de acesso providas no experimento ALICE. Isto irá requerer algumas mudanças na estrutura do ALICE. Com isso estuda-se mudar a posição do FoCal-E para uma distância de 8 m (considerações práticas favorecem a instalação do FoCal-H a $8m$). Este deverá ser construído como um calorímetro hadrônico segmentado convencional, com um tamanho similar ao FoCal-E.

4.2 Motivação: Função de Distribuição de Pártons e o *Color Glass Condensate*

O objetivo do FoCal é o estudo das funções de distribuições de pártons (*Parton Distribctions Functions* - PDF) no regime de pequeno-x (pequeno Bjorken-x e baixo Q^2), em que é esperado que as PDFs evoluam não linearmente devido a alta densidade de glúons, fenômeno conhecido como saturação de glúons (*gluon saturation*). O conhecimento das funções de distribuições dos pártons (*Parton Distribuction Function* - PDF) é importante para os estudo do espalhamento a altas energias. As PDFs para prótons são relativamente bem conhecidas, mas os núcleos não podem ser tratados como uma mera superposição de PDFs de prótons e nêutrons. Devido a isso estas PDFs estão sujeitas a incertezas, o que tem reflexo no estudo de colisões a altas energias (p.e produção de partículas). Isto pode ser ilustrado observando o fator de modificação nuclear:

$$R_{AB} = \frac{\frac{1}{p_T}\frac{dN^h_{AB}(b)}{dp_T}}{\langle N_{coll}(b)\rangle \frac{1}{p_T}\frac{dN^h_{pp}}{dp_T}} \tag{4.1}$$

para colisões do núcleo de massa A com um núcleo de massa B a um parâmetro de impacto b para $1/p_T dN^h_{AB}(b)/dp_T$ a produção invariante de hádrons h por evento, para um dada centralidade, e $1/p_T dN^h_{pp}/dp_T$ a produção invariante de hádrons h por evento $p+p$ inelástico, e $\langle N_{coll}(b)\rangle$ o número médio de colisões nucleon-nucleon binárias para uma dada centralidade.

O cálculo deste fator está intimamente ligado à escolha da PDF. Na figura 4.3 pode-se observar que para a escolha de duas PDFs diferentes tem-se dois valores para o fator de modificação nuclear.

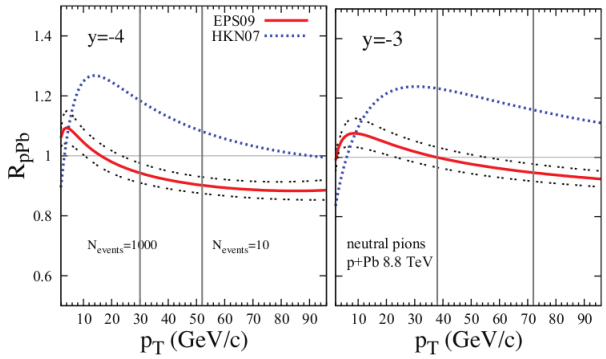


Figura 4.3: Fator de modificação nuclear para píons neutros em colisões $p+Pb$ no LHC utilzando duas escolhas de PDFs. A linha vermelha mostra a EPS09 e a curva azul HKN07(figura retirada de [14]).

A distribuição de pártons tem sido estudada através do espalhamento inelástico profundo (*deep inelastic scattering* - DIS) [13]. Em colisões de íons pesados, a maioria de partículas produzidas tem um momento transversal de alguns GeV. Isto implica que eles são produzidos por pártons que carregam uma pequena fração x do momento longitudinal da partícula. Além do mais, esta fração do momento diminui com o aumento da energia do centro de massa da colisão. Conforme esta energia se torna alta o párton mais relevante para o espalhamento é o glúon, o qual carrega somente uma pequena fração do momento do hádron [15]. Com a diminuição da fração do momento do párton $\left(x = \frac{p_{parton}}{p_{hadron}}\right)$ a densidade de glúons aumenta rapidamente. Quando esta densidade não é tão alta, a evolução deste estado é descrita pelas equações BFKL (Balitski-Fadin-Kuraev-Lipatov) [16][17] e DGALP (Dokshitzer-Gribov-Lipatov-Altareli-Parisi) [18][19]. O aumento da densidade de glúons é balanceado pelo processo de fusão de glúons. Estes efeitos não lineares levam a uma saturação de glúons na região de pequeno-x (*small-x*)[20] [21]. Sabendo que o tamanho do núcleo varia pouco com o aumento de energia, surge a questão de como se adicionar glúons ao sistema. Inicialmente adiciona-se glúons de tamanho fixo. A uma dada energia estes terão preenchido todo o hádron. Interações repulsivas começam a se tornar relevantes e o processo de se acrescentar glúons para. Para continuar aumentando o

número de glúons é necessário que estes novos tenham um "tamanho" menor, ou seja estas partículas tem um x de Bjorken menor. Fazendo uma analogia com esferas rígidas, para se adicionar mais num determinado lugar é necessário que estas novas esferas tenham uma dimensão menor. Este processo continua sem interrupção, sempre adicionando glúons de "tamanhos" menores, conforme aumenta-se a energia[22].

Experimentalmente espera-se que os efeitos desta saturação na região de pequeno-x se revelem em medidas de produção de partículas dianteiras com o aumento da energia do centro de massa da colisão. Também, este estado tem um importante papel na definição das condições iniciais para qualquer interação hadrônica de alta energia e na redução das incertezas nas PDFs dos núcleos. A teoria que descreve os processo desta região é o *Color Glass Consdensate* (CGC). Este é um estado inicial em colisões de hádrons altas energias que descreve a distribuição de quarks e glúons medidas em DIS e está associado a uma alta densidade de glúons. Logo após a colisão entre hádrons, a natureza da matéria muda sua estrutura, dando origem ao Glasma. Quando este decai é formado o Plasma de Quarks e Glúons (*Quark Gluon Plasma* - QGP) que por sua vez decai em hádrons que são medidos nos detectores[22].

Com isso, o CGC permite uma descrição de primeiro princípios dos estágios iniciais das colisões nucleares de alta energia, que envolvem glúons em regime saturados. Ela descreve corretamente a densidade de energia liberada em tais colisões, como também característcas importantes das correlações observadas nos estados finais. Mas mais fundamental, a CGC, dá uma descrição, baseada na QCD, dos momento iniciais da colisão de íons pesados, e é também o arcabouço a ser usado para justificar a aplicabilidade da hidrodinâmica.

4.3 O que vai medir?

Como descrito na seção anterior, os efeitos de saturação favorecem a produção de partículas na região dianteira conforme aumentá-se a energia do centro de massa. Com isso o objetivo primário do FoCal será fornecer medidas precisas de fótons diretos, jatos, e medidas de coincidências gamma-jato e jato-jato em colisões próton-próton e próton-núcleo que irão limitar as PDFs na região de pequenos-x em colisões de íons pesados. Este trabalho se ocupou somente da análise de fótons diretos.

Os fótons diretos são divididos em duas categorias: fótons térmicos e *prompt* γ (γ diretos). Fótons térmicos carregam informação de temperatura do meio em expansão. Já os *prompt* γ são definidos como fótons provenientes de um espalhamento Compton duro ou aniquilação de quarks:

$$\text{Compton:} \quad g + q \rightarrow \gamma + q$$
$$\text{Aniquilação:} \quad q + \overline{q} \rightarrow \gamma + g$$

A medida destes "tipos" de fótons é importante pelos motivos abaixo:

- O estudo de *prompt* γ é importante para a compreensão dos processos que ocorrem no plasma de quarks e glúon (*quark-gluon plasma* - QGP).
- A partir de fótons térmicos é possível entender as propriedades térmicas da fase inicial da reação.
- Como o fóton é uma partícula que interage muito pouco com as outras partículas presente no meio formado (CGC e QGP) ela é uma ponta de prova valiosa para observar as propriedades deste meio.

A identificação destes fótons primários é dificultada devido ao grande fundo produzido por fótons oriundos de decaimentos. Por isso se torna necessária uma medida precisa dos mésons neutros produzidos e um grande poder de discriminação entre *clusters* de *single* fóton (proveniente de *prompt* γ) e π^0. Isto é um grande desafio para momentos transversais altos, onde os *clusters* de π^0 começam a se fundir. Uma maneira seria utilizar o cálculo da massa invariante para determinar se dois chuveiros são provenientes do decaimento do píon. Ela é dada por:

$$m_\pi^2 = 2p_{\gamma_1}p_{\gamma_2}(1 - cos\phi) \tag{4.2}$$

em que p_γ é o momento do fóton do decaimento e ϕ é o ângulo entre as duas partículas filhas.

Tomando por exemplo a figura 4.4 pode-se observar os locais acionados no detector por fótons provenientes de um decaimento $K_L \rightarrow \pi^0 + \pi^0 \rightarrow \gamma + \gamma + \gamma + \gamma$. As regiões acionadas estão sombreadas. Cada conjunto de áreas sombreadas é um *cluster* (regiões dentro do círculo azul). O ponto onde o K_L decaiu está marcado como o losango branco de onde saem linhas tracejadas indicando a trajetória dos fótons até o calorímetro [9].

Agora, supondo que este calorímetro esteja na posição frontal, quando este píon tem baixa energia os fótons detectados estão bem distantes um do outro, sendo possível sua identificação (como mostrado na figura 4.4). Quando a energia deste píon é alta, os fótons do decaimento estarão muito próximos, sendo detectados como um único fóton. Como é importante a identificação de fótons diretos é interessante criar um método que pode separar fótons provenientes de decaimento dos que são diretos.

Uma das maneiras seria diminuir o tamanho do sensor, mas isto não pode ser realizado de maneira arbitrária, devido a limitações tecnológicas e orçamentárias. Neste ponto

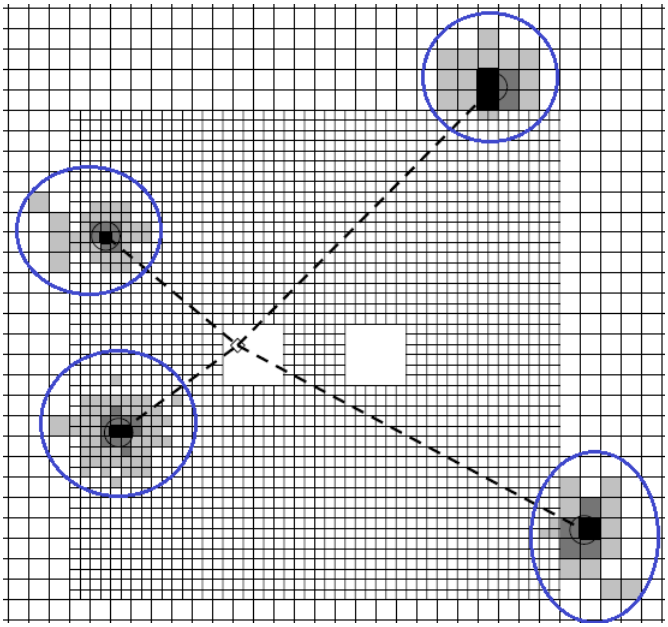


Figura 4.4: Uma visão transversal dos *clusters* no calorímetro KTEV para um evento de decaimento $K_L \rightarrow \pi^0 + \pi^0$. O calorímetro tem uma seção de choque de $5 \times 5cm^2$ ($2.5 \times 2.5cm^2$) na região externa (interna). Retirado e adaptado de [9].

entra em cena métodos computacionais de mineração de dados. Dentre os métodos de mineração, optou-se pelos de aprendizado de máquina. Ainda dentro deste escolheu-se os de aprendizado não supervisionado, pois o algoritmo deve encontrar os *clusters* sem ajuda de um modelo externo. Se fosse escolhido o supervisionado seria necessário que houvesse um banco com todas as formas possíveis de resultados de colisões para que ele pudesse classificar um novo resultado baseado neste banco de dados. Assim, sendo isto inviável, este método não poderia ser escolhido.

As simulações e análises para o FoCal estão sendo realizadas utilizando dois códigos: GEANT 4[23] e FoCal. Para este trabalho foi utilizado somente o último. O FoCal foi desenvolvido por Taku Gunji (Center for Nuclear Study, University of Tokyo)[13] exclusivamente para a simulação e análise deste calorímetro. Este é um código desenvolvido em C++ que utiliza toda a estrutura do existente do AliROOT[24], mas com a adição de bibliotecas específicas para o FoCal. Este pacote será chamado neste trabalho de FoCal Framework.

Capítulo 5

Clusterização

O grupo responsável pelo desenvolvimento do calorímetro percebeu a necessidade de estudar um algoritmo de clusterização para a separação dos sinais deixados por γ proveniente de π^0 e diretos. A análise por *cluster* agrupa objetos (dados) baseados somente na informação contida no conjunto de dados dos mesmos. O seu principal objetivo é que os objetos dentro de um grupo sejam similares (ou relacionados) entre si e ao mesmo tempo diferentes dos objetos de outros grupos. Ou seja, ela tenta encontrar estruturas em conjunto de dados e é por isso exploratório por natureza. Quanto maior a similaridade entre os elementos de um grupo e maior as suas diferenças entre os elementos de outro grupo, mais distintos são os clusters [25][26]. Mas como definir um "bom" cluster? Na verdade não existe um único critério para definir o melhor cluster. Isto depende do objetivo final da análise, e o usuário é que devem ditar este critério para que o resultado final tenha a informação que ele precisa.

A análise por *cluster* está relacionada a outras técnicas de classificação que são usadas para dividir os dados em grupos. Com isso a clusterização pode ser entendida como uma forma de classificação em que se qualifica cada dado em uma classe. Isto é realizado somente com o que esta disponível no conjunto de dados. A ausência de categorias bem definidas neste conjunto de dados diferencia a clusterização (aprendizado não supervisionado) da análise por classificação ou discriminação (aprendizado supervisionado). Um dos algoritmos mais famosos e simples é o K-means (publicado em 1950)[27].

O aumento no volume e tipo de dados requer técnicas avançadas para automaticamente compreender, processar e resumir o conjunto de dados. As técnicas para mineração de dados podem ser classificadas em dois grandes grupos [27]: *exploratórias* ou descritivas, significando que o investigador não tem modelos pré-definidos ou hipóteses, mas deseja entender as características ou estruturas em um conjunto de dados de várias dimensões; o outro grupo é o de *confirmação* ou de inferência, significando que o investigador quer confirmar a validade de uma hipótese/teoria ou um conjunto de afirmações dado um certo conjunto de dados.

Em reconhecimento de padrões (*pattern recognition*), a análise de dados esta interessada em modelos preditivos: dado um conjunto de dados para treino (conhecido), queremos que seja predito o comportamento de um conjunto de dados desconhecido. Este trabalho é também conhecido como aprendizado. Frequentemente uma clara distinção é feita entre problemas de aprendizados que são supervisionados (classificação) ou não-supervisionado (clusterização). O primeiro envolve somente dados já classificados (padrões com categorias já conhecidas) e o segundo com dados completamente desconhecidos. Clusterização é um problema muito mais difícil e desafiante que classificação. Atualmente há um grande interesse em modelos de mineração híbridos, chamados de aprendizado semi-supervisionado. Neste modelo, as classificações estão disponíveis para uma pequena parte dos dados. O dados sem classificação em vez de serem descartados também são usados no processo de aprendizado. No aprendizado semi-supervisionado em vez de se especificar as classificações, são utilizados comparações entre objetos[27].

5.1 O que é o aprendizado de máquina?

Aprendizado, como a inteligência, abrange vários tipos de processos que são difíceis de se definir precisamente. Uma definição de dicionário incluem frases como " Ficar sabendo, reter na memória, tomar conhecimento de, aprender por experiência própria" e "modificação de comportamento por experiência". Zoologistas e psicólogos estudam o aprendizado em animais e humanos. Há vários paralelos entre animais e aprendizado em máquinas. Certamente, muitas técnicas em aprendizado de máquinas derivam de esforços de psicólogos para tornar suas teorias de aprendizados em humanos e animais mais precisa através de modelos computacionais. Parece que algumas técnicas e conceitos explorados por pesquisadores em aprendizado de máquinas podem iluminar certos aspectos do aprendizado biológico[28].

Em relação as máquinas, pode-se dizer, aproximadamente, que uma máquina aprende quando ela muda a sua estrutura, programa, ou dados (baseado em suas entradas ou resposta a uma informação externa) de uma maneira em que é esperado que sua performance futura melhore. Por exemplo, quando a performance de uma máquina de reconhecimento de voz melhora após ouvir várias amostras de falas de pessoas, pode-se dizer, com certo conforto, que a maquina aprendeu[28].

O aprendizado de máquina usualmente se refere a mudanças em sistemas que executam tarefas associadas a inteligência artificial (IA). Tais tarefas envolvem reconhecimento, diagnóstico, planejamento, controle robótico, predição, etc. As mudanças podem ser aprimoramentos de sistemas implantados ou a síntese, do início, de novos sistemas[28].

Pode-se perguntar "Por que máquinas deveriam aprender? Por que não projetar máquinas para fazer o que é desejado em primeiro lugar?" Há várias razões do porque o

aprendizado de máquina é importante. Claro, foi mencionado que a conquista do aprendizado em máquinas pode ajudar a compreensão de como animais e humano aprendem. Mas há importante questões de engenharia também. Algumas são[28]:

- Algumas tarefas não podem ser bem definidas por exemplos; isto é, pode-se ser capaz de especificar a entrada/saída de pares, mas não uma relação precisa entre eles. é conveniente que as máquinas sejam capazes de se ajustarem a estrutura interna para produzir saídas corretas para um grande número de amostras e assim restringir adequadamente a função que rege a relação da entrada e saída para se aproximar a relação implícita entre os exemplos.

- é possível que escondidas entre uma grande quantidade de dados estejam importantes relações e correlações. Métodos de aprendizado de máquina podem frequentemente ser usados para extrair essa relações.

- Os planejadores geralmente produzem máquinas que não trabalham como o desejado nos ambientes em que são usados. De fato, certas características do ambiente de trabalho podem não ser conhecidas no momento do desenvolvimento da máquina. Os métodos de aprendizado de máquina podem ser usados no momento para melhorar *design* das máquinas.

- A quantidade de conhecimento disponível para determinada tarefa pode ser muito grande para os seres humanos processarem. As máquinas que aprendem este conhecimento podem ser capazes de capturar mais do que os humanos queiram escrever.

- O ambiente muda com o tempo. Máquinas que se a adaptam a mudanças no ambiente diminuiriam a constante necessidade para *redesign*.

- Novos conhecimentos sobre a tarefa estão constantemente sendo descoberto por humanos. O vocabulário muda. Há sempre algo novo ocorrendo no mundo. Modificar continuamente sistemas de IA para adaptar novos conhecimentos não é prático, mas métodos de aprendizado de máquinas podem se capazes de se manter atualizadas.

5.1.1 Tipos de aprendizado

Há duas possibilidades com a qual deseja-se aprender uma função (objeto que cria uma relação entre a entrada e a saída do conjunto de dados.). No chamado aprendizado supervisionado, sabe-se (algumas vezes aproximadamente) os valores da função para um conjunto de m de amostras num conjunto de treino. Assume-se que se pode-se encontrar uma hipótese que se aproxima da função desejada para os membros do conjunto de treino, então esta hipótese será um bom "chute", especialmente se o conjunto de treino é grande.

O ajuste de curva é um exemplo simples de aprendizado supervisionado de uma função [28].

Na outra possibilidade chamada de aprendizado não supervisionado, tem-se simplesmente um conjunto de dados para treino e valores de uma função para eles. O problema neste caso, tipicamente, é particionar o conjunto de treino em subconjuntos de uma maneira apropriada (pode-se ainda tratar como se fosse um de aprendizado de uma função: o valor da função é o nome do subconjunto o qual a entrada pertence). Métodos de aprendizado não supervisionado têm aplicações em problemas de taxonomia no quais é desejável inventar maneiras para classificar os dados em categorias que façam sentido.

No aprendizado supervisionado tentá-se fazer uma previsão baseada em um conjunto de dados de entrada. As previsões são baseadas numa amostra de treino de casos previamente resolvidos, em que os valores de todas as variáveis são conhecidos. Isto é chamado de "aprendendo com um professor". Dentro desta metáfora o "estudante" apresenta uma resposta (previsão/saída) para cada valor de entrada e o "professor" diz se a resposta está correta ou não[29].

Para cada exemplo o objetivo é usar as entradas para predizer os valores da saída. Este execício é chamado de aprendizado supervisionado. A entrada é frequentemente chamada de preditor, e mais classicamente de variável independente. Na literatura de reconhecimento de padrões o termo características (*features*) é preferido. As saídas são chamadas de respostas ou classicamente de variável dependente[29].

O objetivo do aprendizado não supervisionado, ou "aprendizado sem um professor", é inferir a partir dos dados, utilizando suas próprias propriedades, as relações de entrada e saída sem um "professor" para mostrar as respostas corretas ou o grau de erro para cada observação. A dimensão do conjunto de dados é maior que o do aprendizado supervisionado, e as propriedades de interesse são muito mais complicadas do que uma simples estimação local[29].

Em aprendizado supervisionado há uma clara medida do sucesso ou a falta dele, que pode ser usada para comparar a efetividade de diferentes métodos em várias situações. No contexo do aprendizado não supervisionado, não há medida direta do sucesso. é difícil determinar a validade da inferência obtida pela saída da maioria dos algoritmos de aprendizado não supervisionado. Com isso deve-se recorrer a argumento heurístico, não somente para motivação dos algoritmos, como também para o julgamento da qualidade dos resultados. Esta situação desconfortável tem levado a proliferação de vários métodos, já que sua efetividade é matéria de opinião e não pode ser verificado diretamente[29].

Considere os vários conjuntos de pontos num espaço bidimensional ilustrado na figura 5.1. O primeiro conjunto (a) parece que pode ser dividido em duas classes naturalmente, enquanto o segundo (b) parece difícil de se dividir, e o terceiro (c) é problemático [28].

Aprendizado não supervisionado usa procedimentos que tentam encontrar divisões

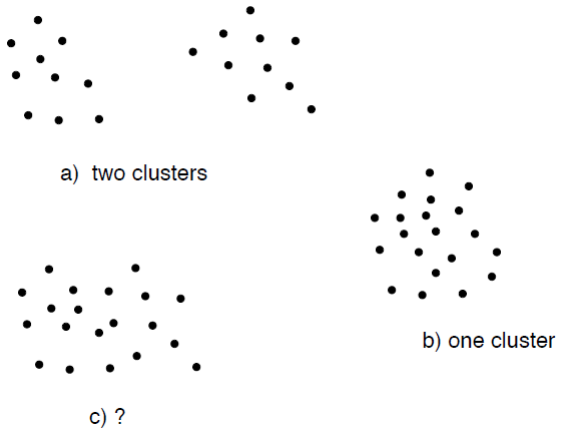


Figura 5.1: Padrões não identificados [28].

naturais da amostra. Há dois estágios [28]:

- Dado um conjunto de dados de uma amostra de treino sem nenhuma identificação, separá-la em vários subconjuntos chamados de clusters.
- Desenvolver um classificador baseado nas identificações dada ao conjunto de treino pelo particionador.

Um tipo de aprendizado não supervisionado envolve encontrar hierarquias da partição ou cluster de clusters. Uma partição hierárquica é uma na qual o conjunto é dividido em subconjuntos mutualmente exclusivo e exaustivo e estes subconjuntos dividido em outros subconjuntos mutualmente exclusivos e exaustivos[28].

5.1.2 Similaridade e Tipos de dados

A similaridade[1](ou dissimilaridade) é de fundamental importância para a definição de um cluster. Esta medida entre dois objetos de um mesmo conjunto de dados é essencial para a maioria dos procedimentos de clusterização. Devido à variedade de tipos dados e escalas, a medida (ou medidas) de similaridade deve ser escolhida cuidadosamente sendo comum calculá-la usando a medida de distância.

A métrica mais popular para características contínuas é a distância Euclidiana. Ela tem um apelo intuitivo já que é comumente usada para calcular a proximidade de objetos em duas ou três dimensões. Ela funciona bem quando os dados apresentam clusters compactos ou isolados[30].

[1]O termo similaridade deve ser compreendido como uma similaridade matemática, medida de uma maneira bem definida. Em espaços métricos, a similaridade é frequentemente definida como a norma[30].

O cálculo das distâncias entre os objetos com algumas das características não contínuas é problemática, já que diferentes tipos de características não são comparáveis e a noção de proximidade é binária para características nominais[30].

Os tipos de dados que se encontram são:

- Variáveis quantitativas: Medidas deste tipo de variáveis ou atributos são apresentados por valores reais contínuos.
- Variáveis Ordinárias: Os valores deste tipo de variáves são frequentemente representados por inteiros contíguos, e os valores são considerados um conjunto ordenado.
- Variáveis categóricas: Com variáveis categóricas sem ordenação (também chamadas de nominais), o grau de diferença entre pares de valores devem ser explicitamente definidos. Dentro deste tipo existem um outro tipo de variável que é a categórica ordenada, por exemplo pequeno, médio e grande, em que há uma ordem entre osvalores, mas nenhuma noção de métrica apropriada (a diferença entre médio e pequeno não precisa ser a mesma entre grande e médio)[29][30].
- Intervalo de escala: Se a diferença entre dois dados podem se expressos como um número, além dos termos mencionados acima.
- Razão: Este tipo de dado é intervalo de escala, mas o valor zero pode existir.

5.2 Análise por cluster

Análise por cluster, também chamado de segmentação de dados, tem um variedade de objetivos. Todos relacionados ao agrupamento ou segmentação de uma coleção de objetos em subgrupos ou "cluster", tal que os objetos dentro de cada cluster são mais semelhante que outros de clusters diferentes. Um objeto pode ser descrito por um conjunto de medidas, ou por sua relação com outros objetos. Em adição, o objetivo é algumas vezes arranjar os clusters em uma hierarquia natural. Isto envolve agrupamentos sucessivos dos próprios clusters em que a cada nível da hierarquia, clusters dentro de um mesmo grupo são mais similares do que outros de grupos diferentes[29].

Análise por cluster é também usada para formar uma estatística descritiva para verificar se um dado consiste ou não de um conjunto de subgrupos distintos, cada grupo repensentando objetos com propriedades substancialmente diferentes. Este último objetivo requer uma avaliação do grau de diferença entre os objetos alocados para os seus respectivos cluster[29].

A noção de similaridade (ou dissimilaridade) entre objetos sendo agrupados é central na análise de cluster. Um método de clusterização tenta agrupar objetos baseado na

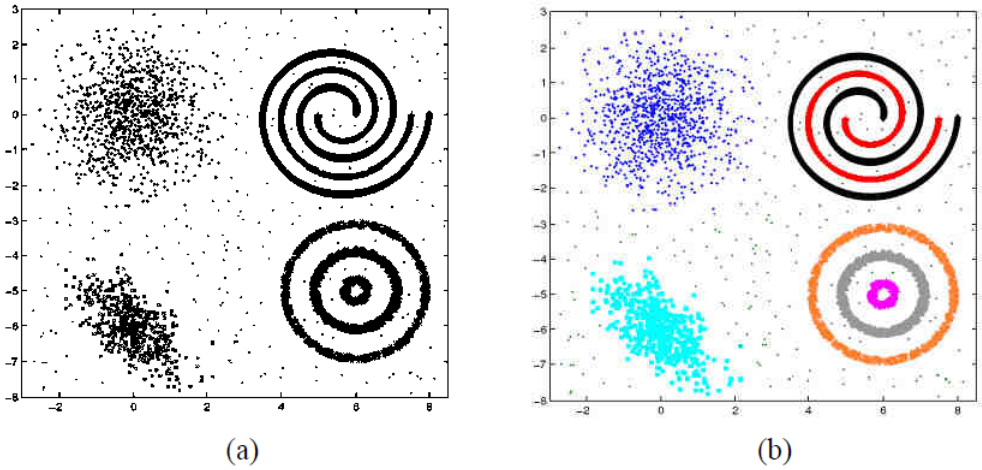


Figura 5.2: Diversidades de clusters. Os setes clusters em (a) (denotados por sete diferentes cores em (b)) diferem em forma, tamanho, e densidade. Embora estes clusters sejam aparentes na análise, nenhum algoritmo de clusterização disponível pode detectar todos eles. Retirado de [27].

definição de similaridade fornecida para ele. Isto pode somente vir do que está sendo analisado. A situação é de alguma maneira similar a especificação da função de perda ou custo em problemas de previsão (aprendizado supervisionado). Há o custo associado com a imprecisão da previsão que depende das considerações fora dos dados[29].

O objetivo da clusterização, também chamado de análise por clusters, é descobrir grupos naturais em um conjunto de dados, pontos ou objetos. Uma definição dada por Webster é a seguinte: uma técnica estatística de classificação para descobrir se indivíduos de uma população se encaixam em diferentes grupos baseados em uma comparação quantitativa de várias características[27].

Uma definição operacional de clusterização é dada a seguir: Dada a representação de n objetos, encontre K grupos baseados em suas semelhanças, sendo que estas semelhanças entre objetos do mesmo grupo sejam grandes e entre grupos diferentes sejam pequenas. Agora surge uma questão, qual é a noção de semelhança? Qual é a definição de cluster a ser tomada? Na figura 5.2 pode-se observar que os clusters podem ser diferentes em termos de forma, tamanho e densidade[27].

Um cluster ideal pode ser definido como um conjunto de pontos que é compacto e isolado. Na realidade um cluster é uma entidade **subjetiva** que está no olho do observador e o seu significado e interpretação requerem conhecimento por parte do observador. Enquanto seres humanos são excelentes para encontrar cluster em duas ou três dimensões, para dimensões maiores é necessário um algoritmo para automatizar o processo. Para este desafio e para o desconhecimento do número de clusters contido num conjunto de dados

um número extremamente grande de algoritmos surgiu e continua a surgir[27].

A análise de dados por cluster tem sido usada para os seguintes propósitos:

- Descobrir estruturas: ganhar informação sobre os dados, gerar hipótese, detectar anomalias e indentificar características salientes.
- Classificação Natural: identificar o grau de similaridade entre formas ou organismos (relações filogenéticas).
- Compressão: método para organizar dados e resumí-los através de clusters.

O desenvolvimento do método de clusterização tem sido um verdadeiro esforço multidisciplinar. Taxonomistas, cientistas sociais, psicólogos, biólogos, estatísticos, matemáticos, engenheiros, cientistas da computação, médicos (pesquisadores), e outros que coletam e processam dados reais têm contribuído para o método da clusterização[27].

Os algoritmos de clusterização podem ser divididos em dois grupos: hierárquicos e de particionamento (*partitional*). Algoritmos de clusterização hieráquicos encontram recursivamente clusters aninhados tanto no modo aglomerativo (começando com cada ponto sendo um cluster e juntando os mais similares sucessivamente até formar uma hierarquia de clusters) como no modo divisivo (começando com todos os pontos sendo um único cluster e separando-os sucessivamente). Algoritmos de particionamento encontram todos os clusters simultaneamente, sem impor uma estrutura hierárquica. A entrada de um algoritmo de hierarquia é uma matriz de similaridade $n \times n$, em que n é o número de objetos a ser separado em clusters. Por outro lado o de particionamento pode usar uma matriz de similaridade $n \times n$ como uma matriz $n \times d$, em que os n objetos estão num espaço de d-dimensões. Os algoritmos hierárquicos mais conhecidos são os ditos *sigle-link* e *complete-link* [25] e o de particionamento é o K-means [25][27].

Cluster podem ser definidos como regiões de alta densidade no espaço das características separados por regiões de baixa densidade. Algoritmos que seguem esta noção de cluster procuram por regiões muito densas neste espaço. Diferentes algoritmos usam diferentes definições de conexões. O algoritmo de Jarvis-Patrick define a similaridade entre um par de pontos como o número de vizinhos que eles compartilham, sendo que estes estão dentro de um raio pré-definido em torno do ponto [31]. DBSCAN [25] procura por regiões densas e estima sua densidade utilzando o método da janela de Parzen [32]. Ambos algoritmos citados dependem de dois fatores: tamanho da vizinhança e número minimo de vizinhos para sua inclusão num cluster.

Em contraste com o grande número de algoritmos utilizados e seus sucessos em várias áreas e aplicações, a clusterizaação continua sendo um problema difícil. Isto pode ser atribuído à falta de clareza na definição de cluster e na definição da similaridade entre os objetos. Os seguintes desafios fundamentais associados com a clusterização, são:

- O que é um cluster?
- Quais características deve-se usar?
- Os dados devem ser normalizados?
- Como definir a comparação entre dois objetos?
- Quantos clusters estão contidos no conjunto de dados?
- Qual método de clusterização deve-se utilizar?
- Os dados contém algum prévio agrupamento?
- Os clusters encontrados são válidos?

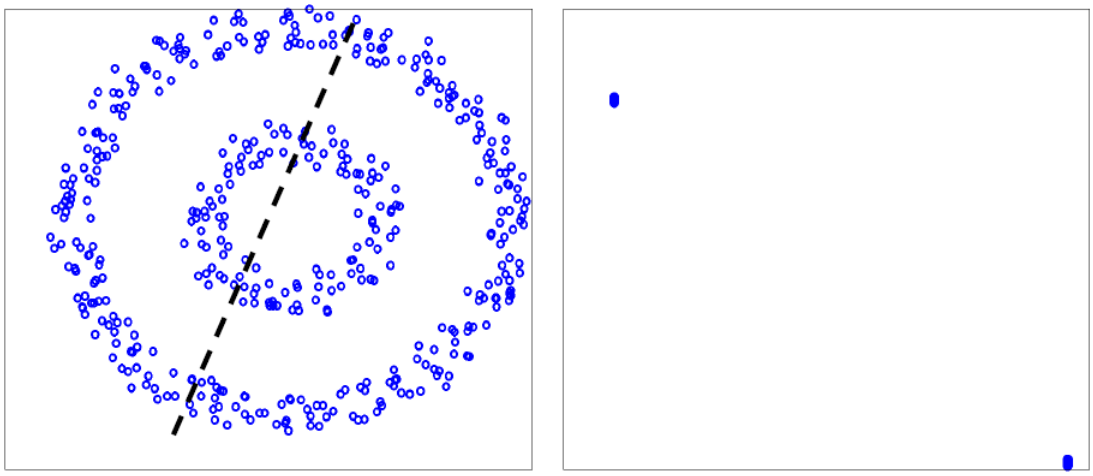
Como os dados estão dispostos (*data representation*) é um dos principais fatores que influenciam na performance do algoritmo de clusterização. Se a representação (escolha das características mais importante para determinar um cluster) for boa, os clusters encontrados provavelmente serão compactos e isolados e até mesmo um algoritmo simples como o K-means irá encontrá-los. Infelizmente não há uma representação universal e a escolha da representação deve ser guiada pelo conhecimento do problema. Na figura 5.3 pode-se observar que um conjunto de dados não é muito bem descrito pelo K-means, mas quando utilizado os dois autovetores da matrix de RBF (*radial basis function*), os clusters se tornam bem separados e facilmente identificados pelo K-means [27]. A representação dos dados está intimamente relacionada ao propósito de agrupamento.

Determinar automaticamente o número de clusters tem sido um dos problemas mais difíceis na clusterização. A maioria dos métodos que determinam automaticamente o número de clusters recaem no problema do modelo a ser utilizado. Usualmente os algoritmos de clusterização rodam com diferentes número de clusters K. O melhor valor de K é escolhido baseado num critério pré-definido[27].

Um ponto importante é a validade do cluster. **Algoritmos de clusterização** tendem a encontrar clusters nos dados independente deles exitirem ou não. Existem três critérios para validar um cluster: interno, externo e relativo. Índices baseados em critérios internos, avaliam o ajuste entre a estrutura imposta pelo cluster e os dados, utilizando somente os dados. Índices baseados em critérios relativos comparam mútiplas estruturas (geradas por diferentes algoritmos) e decidme qual deles é melhor. Índices externos medem a performance comparando a estrutura do cluster com uma definida anteriormente e tida como verdadeira[27].

Algumas das aplicações da análise por *cluster* são:

- *Marketing*: encontrar grupos de consumidores com comportamento semelhante baseado em suas compras passadas;



(a) (b)

Figura 5.3: Importância da representação dos dados. (a) Dois anéis concêntricos que o K-means falha para encontrar os clusters naturais. A linha tracejada mostra a fronteira entre os dois clusters encontrados pelo K-means. (b) Uma nova representação dos dados de (a) utilizando RBF. Neste caso o K-means pode facilmente encontrar os dois clusters.

- Biologia: classificação taxonômica;
- Bibliotecas: organização dos livros;
- Planejamento de cidades: identificar grupos de casas de acordo com seu valor, tipo e localização geográfica;
- Terremotos: identificação de zonas perigosas baseado em conjuntos (*cluster*) de epicentros observados.

Os principais requerimentos para um algoritmo de clusterização são:

- Lidar com diferentes tipos de dados;
- Encontrar *cluster* com formas arbitrárias;
- Necessidade de mínimo conhecimento do objeto de estudo para determinar os parâmetros de entrada;
- Habilidade para lidar com *outliers* e ruídos;
- Insensibilidade à ordem dos dados de entrada;
- Trabalhar com vários parâmetros;
- Usabilidade e interpretabilidade.

Abaixo segue uma lista de alguns problemas que podem ser encontrados nos algoritmos:

- As técnicas atuais de clusterização não lidam com todas as necessidades adequadamente;
- Com o aumento dos dados e parâmetros aumenta-se o tempo para concluir a tarefa;
- A efetividade do método depende da definição da dissimilaridade;
- Complicações para definir a melhor dissimilaridade a ser utilizada;
- Os resultados da clusterização podem ter várias interpretações.

5.3 Técnicas de Clusterização

Nesta seção serão apresentados alguns algoritmos mais conhecidos para análise através da clusterização.

5.3.1 Algoritmos de Particionamento

Um algoritmo de particionamento obtém uma única partição dos dados ao invés de uma estrutura de clusters. Um problema que acompanha o uso destes algoritmos é a escolha do número de clusters finais. Técnicas de particionamento usualmente produzem clusters pela otimização de uma função definida localmente (num subconjunto dos dados) ou globalmente (em todos os dados). Executar todas as opções possíveis para chegar a um critério ótimo é computacionalmente proibitivo. Na prática, por isso, o algoritmo é executado várias vezes com diferentes configurações iniciais, e a melhor configuração é obtida a partir de todos os resultados destes testes[30].

Outra variação é permitir a separação e a união dos clusters. Tipicamente, um cluster é separado quando sua variância está acima de um limiar especificado, e dois clusters são unidos quando a distâncias de suas centróides estão abaixo de um valor pré-estabelecido. Usando esta variante é possível obter a partição ótima começando de condição inicial, desde que haja valores de limiares estabelecidos[30].

K-Means O algoritmo *K-means* é um dos métodos de clusterização mais populares, por isso irá-se eplicá-lo em maiores detalhes. Ele é usado em situações nas quais todas as variáveis são do tipo quantitativo e usa a distância Euclidiana

$$d(x_i, x_{i'}) = \sum_{j=1}^{p} (x_{ij}, x_{i'j'})^2 = ||x_i, x_{i'}||^2 \tag{5.1}$$

como medida da dissimilaridade. Ele começa com partições randomicamente distribuídas. Ele fica redistribuindo os dados em clusters baseado na similaridade entre o dado e o centro do cluster até que o critério de convergência seja atingindo. Este algoritmo é popular porque é de fácil implementação e seu tempo de execução cresce linearmente com a quantidade de dados. Uma de suas desvantagens é que ele é sensível as condições iniciais e pode convergir para um mínimo local se a condição não foi corretamente escolhida[30][29].

Escolha do número de clusters A escolha do número de clusters K depende do objetivo. Para segmentação de dados, K é usualmente definido como parte do problema. Por exemplo, uma companhia pode empregar K vendedores, e o objetivo é particionar uma base de dados de clientes em K segmentos, um para cada vendedor, tal que os clientes atribuídos a cada um sejam os mais semelhante possível. Frequentemente, contudo, a análise por cluster é usada para prover uma estatística descritiva para determinar a extensão de qual observações caem em grupos naturais distintos. Aqui o número de tais grupos K^* não é conhecido, e que este juntamente com os próprios grupos, seja estimado dos dados[29].

5.3.2 Algoritmo de Clusterização Hierárquica - Clusterização Aglomerativa

Suponha que se tenha um conjunto de treino com padrões desconhecidos. Pode-se formar uma classificação hierárquica dos padrões por um método aglomerativo (a descrição deste método é baseado num manuscrito não publicado de Pat Langley)[28].

Primeiramente calcula-se a distância euclidiana entre todos os elementos do conjunto. Suponha que a menor distância seja entre os elementos X_i e X_j. Estes dois são unidos em um cluster C e no conjunto de treino serão substituídos por este cluster. Após isso todas as distâncias entre os elementos são recalculadas. Se a menor distância ocorrer entre outros dois elementos, estes dois serão substituídos por um outro clusters C^* e realiza-se o procedimento anterior. Se a distância menor é entre um objeto e um cluster, os dois são unidos e substituído pela média entre eles (com seus pesos apropriados). Quando a menor distância é entre dois clusters ambos são unidos e substituídos por sua média (com seus pesos apropriados) e continua-se o processo. Já que o número de pontos é reduzido, no final têm-se uma árvore de clusters em que na base estão todos os pontos do conjunto original[28].

Um exemplo deste método pode ser mostrado na figura 5.4. Os números associados com cada cluster indica a ordem em que foram formados[28].

Há duas variantes do algoritmo mencionado acima que são mais usadas: *single-link* e *complete-link*. Este dois algoritmos diferem na maneira em que eles caracterizam a

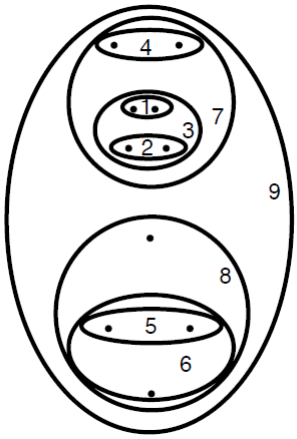


Figura 5.4: Cluster aglomerativo[28].

similaridade entre um par de clusters. Suponha que G e H representam dois grupos. No método de *single-link*, a distância é a mínima das distâncias entre todos os pares de dados dentro de dois clusters distintos. Isto irá levar clusters unidos por uma série de observações intermediariamente próximas. Este método também é chamado frequentemente como a técnica do vizinho mais próxima (*nearest-neighbor*). No método de *complete-link*, a distâcia entre dois clusters é a máxima dentre todas as distâncias entre dois dados no interior de dois clusters distintos. Isto tenderá a produzir clustes compactos com diâmetros pequenos[30][29].

5.3.3 Clusterização Hieráquica - Clusterização Divisiva

Algoritmos de clusterização divisiva começam com todo conjunto de dados como um único cluster, e recursivamente divide este em outros clusters e assim sucessivamente na direção de cima para baixo. Esta abordagem não tem sido estudada tanto como os aglomerativos. Dentro da clusterização, uma vantagem potencial dos métodos divisivos sobre os aglomerativos pode ocorrer quando se está interessado em dividir os dados em relativamente poucos clusters[29].

O modelo da divisão pode ser empregado por aplicar recursivamente qualquer um dos métodos combinatórios tais como K-means ou K-medoids, com $K = 2$ para realizar a separação em cada iteração. Contudo, tal aboradagem dependeria da configuração inicial especificada a cada passo[29].

Um exemplo de algoritmo de divisão é o que foi proposto por Macnaughton Smith. Ele começa por colocar todos os dados num único cluster G. Então escolhe-se o dado em que a similaridade média é a maior de todas. Esta observação forma o primeiro membro do segundo cluster H. A cada passo sucessivo aquela observação em G que tenha

a similaridade maior é tranferida para H. Isto continua até que as diferenças entre as médias seja negativa, isto é, não há nenhuma observação em que, em média, é próxima àquelas em H. O resultado é uma separação do cluster original em dois clusters filhos, as observações transferidas para H, e as que ficaram em G. Estes dois clusters representam o segundo nível da hierarquia. Cada nível sucessivo é produzido por este processo de separação de um cluster do nível anterior. Alguns autores sugerem que deva-se escolher o cluster com o maior diâmetro para serparação. Este processo continua até que os cluster se tornem *singletons* (clusters formados por um único elemento) ou todos os membros tenha cada um similaridade zero entre eles[29].

5.3.4 Clusterização por Fuzzy

Clusters podem ser formalmente vistos como subconjuntos do conjunto de dados. Uma possível classificação para os métodos de clusterização depende se os subconjuntos são fuzzy ("impreciso") ou *hard* ("rígido"). Clusterização *hard* são baseados na teoria de conjunto clássica, e requer que um objeto pertença ou não a um cluster. Clusterização *hard* significa que um conjunto de dados é dividido num número especificado de subconjuntos mutuamente exclusivos. Os métodos de clusterização fuzzy permitem que objetos pertençam a vários clusters simultaneamente, com diferentes graus de filiação a estes clusters. Em várias situações reais, a clusterização fuzzy é mais natural que uma clusterização *hard*, já que os objetos na fronteira não são forçados a pertencerem a um determinado cluster, mas a eles são dados um grau de filiação a estes num valor entre 0 e 1[30].

5.4 Descrição do algoritmo utilizado neste trabalho

O algoritmo desenvolvido neste trabalho se encaixa no tipo de particionamento. Será descrito o algoritmo de clusterização implementado no FoCal Framework, pois foi por este que se começou o desenvolvimento do algoritmo de clusterização deste trabalho. O desenvolvimento destes algoritmo será apresentado em três partes. Elas simplesmente significam que quando utilizei este algoritmo eles estavam no estágio que serão apresentados abaixo, todos são uma evolução do algoritmo 1 (algoritmo no estágio de desenvolvimento 1).

5.4.1 Algoritmo 1

Inicialmente determina-se qual será torre[2] central do futuro clusters. Isto executá-se comparando a energia de cada pad com o seu respectivo vizinho. Com este pad,

[2]unidade sensível do calorímetro. Tem o tamanho do pad (observando o detector transversalmente) e o comprimento do segmento. O segmento é o conjunto de *layers* que são lidos em conjunto.

determina-se uma matriz de pad 3×3 ao redor dele. A energia do pad central é somada à dos vizinhos. Suas posições também são somadas, mas são ponderadas pela energia. Depois disso estas variáveis, que definem um cluster (objeto), são armazenadas.

Inicia-se uma varredura nestas variáveis. Primeiramente calcula-se a distância entre os clusters do primeiro segmento e os clusters do segundo utilizando a equação abaixo,

$$dx = x_1 - \frac{x_0}{z_0} * z_1 \tag{5.2}$$

$$dy = y_1 - \frac{y_0}{z_0} * z_1 \tag{5.3}$$

em que x_0, y_0 e z_0 correspondem as posições x, y e z, respectivamente, do objeto do primeiro segmento e x_1, y_1 e z_1 correspondem as posiçẽs x, y e z, respectivamente, do objeto do segundo segmento. Se dx e dy estão dentro do intervalo]-5,5[, calcula-se a distância do objeto do segundo segmento com do terceiro utilizando a equação abaixo

$$dx = x_2 - \frac{x_1}{z_1} * z_2 \tag{5.4}$$

$$dy = y_2 - \frac{y_1}{z_1} * z_2 \tag{5.5}$$

em que x_2, y_2 e z_2 correspondem as posiçẽs x, y e z, respectivamente, do objeto do terceiro segmento. Se este resultado também estiver dentro do intervalo]-5,5[esta variáveis(correspondente a união dos três segmentos) também são armazenadas sendo elas: a soma da energia e as posiçẽs x, y e z ponderadas pela energia. No caso de houver detectores de *strip* na geometria adotada (neste trabalho não foi utilizada nenhuma geometria com *strip*), ocorre uma última associação entre os cluster obtidos nos pad e *strip*. São tomadas as posições x, y, z e energia de cada um e calculá-se as distâncias abaixo:

$$dx_f = x_{pad} - \frac{x_{strip}}{z_{strip}} * z_{pad} \tag{5.6}$$

$$dy_f = y_{pad} - \frac{y_{strip}}{z_{strip}} * z_{pad} \tag{5.7}$$

$$\tag{5.8}$$

Após isto se eliminam as falsas associações entre os clusters dos pads e *strip*, calculá-se novamente dx_f e dy_f e por último, calculá-se:

$$d_f = \sqrt{dx_f^2 + dy_f^2} \tag{5.9}$$

Sendo que $d_f < 5$ para se unir estes clusters.

As equações 5.2 e 5.3 foram deduzidas da seguinte maneira (esta dedução também se encaixa para as equações 5.6 e 5.7). Observando a figura 5.5 tem-se uma representação

pictórica de um calorímetro. Nele assume-se que a partícula que irá dar origem ao chuveiro não tem sua trajetória desviada. Neste exemplo a partícula desenvolve seu chuveiro e em duas camadas são encontrados um cluster A e B respectivamente. Uma questão que se levanta é, estes dois clusters pertencem à mesma partícula? Para resolver isso se procedeu da seguinte maneira: Traça-se uma reta da origem (Ponto de interação) até o cluster encontrado na primeira camada (A). Com isso tem-se que:

$$Y_0 = aZ_0 \Rightarrow a = \frac{Y_0}{Z_0} \Rightarrow Y = \frac{Y_0}{Z_0}Z \tag{5.10}$$

A posição Z é a posição do centro do *layer*. Agora, calcula-se a posição do cluster A no próximo segmento para se saber qual seria a posição da partícula incidente caso não houvesse nada no caminho dela. Assim tem-se a posição do cluster B (Y_1) e a do cluster A(Y_1') projetada,

$$Y_1' = \frac{Y_0}{Z_0}Z_1 \tag{5.11}$$

subtraindo as equações 5.10 e 5.11 tem-se:

$$dy = Y_1 - \frac{Y_0}{Z_0}Z_1. \tag{5.12}$$

Para dx realiza-se o mesmo procedimento. Como critério para saber se os clusters A e B pertencem à mesma partícula (ou seja são do mesmo chuveiro) testa-se para $d < 5$, se isto for verdade então unem-se estes dois clusters (soma da energia e as posições x, y e z ponderadas pela energia) obtêndo-se assim o cluster final. Os resultados com este algoritmo estão apresentados na seção 6.1.

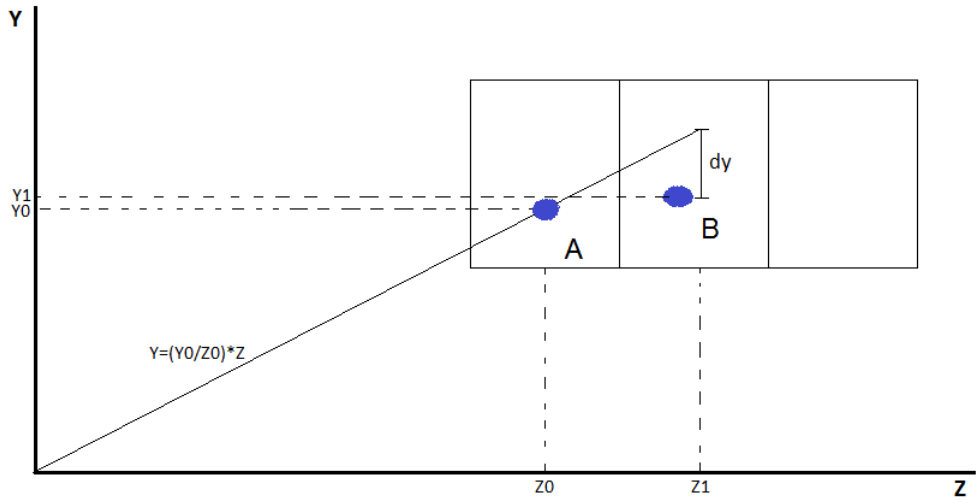


Figura 5.5: Esquema de um calorímetro atravessado por uma partícula que tenha deixado, provavelmente dois clusters.

5.4.2 Algoritmo 2

Após estes resultados da seção 6.1 notou-se que o este algoritmo era muito rígido. Caso fosse necessária uma mudança radical na geometria utilizada para análise, precisaria-se executar uma grande reestruturação no algoritmo. Com isto pensou-se várias modificações para que ele se ajustasse às novas geometrias que poderiam surgir e para melhorar seu desempenho. Assim, com a evolução da geometria, novas unidades básicas para o calorímetro foram surgindo. As definições destas unidades estão descritas abaixo:

- Pad: sensor do calorímetro. Pode ser *pads* ou pixels (figura 5.6).
- Segmento: seqüência de *layers* que são lidos em conjunto numa medida (figura 5.7).
- Célula: unidade virtual que tem o tamanho do Pad (plano XY) e o comprimento do segmento (figura 5.6 e figura 5.7).

No estágio de desenvolvimento anterior, apesar do algoritmo também dividir o detector em segmentos (a única divisão era em três), podía-se acessar os dados obtidos em cada *layer*. Nesta modificação isso não é mais possível, sendo somente acessadas as informações obtidas pelos segmentos. Isto foi realizado pois no detector real isto também ocorreá. Abaixo segue uma descrição do algoritmo:

1 Usando um vetor contendo todas as células que tiveram sinal, o algoritmo procura por aquelas com energia acima de um limiar determinado pelo usuário e cria um novo vetor com estas;

2 Depois disto, o algoritmo procura pela célula com a energia maior dentro de um raio (em cada segmento) e chama esta de centro do sub-cluster (cluster encontrado em cada segmento);

3 Uma vez encontrados tais centros, ele determina que todas as células localizadas a 5 cm de distância em x e y do centro são vizinhas;

4 A energia da célula central é somada à energia das células vizinhas, com isto formando um sub-cluster. Neste ponto, uma célula vizinha pode fazer parte de mais de um sub-cluster;

5 Se o usuário desejar a informação sobre os sub-cluster, o programa terminará aqui;

6 Caso o usuário queira a informação do cluster de todo o detector, a combinação entre os sub-clusters será realizada calculando a distância entre os sub-clusters de

cada segmento utilzando a equação abaixo,

$$dx = x_1 - \frac{x_0}{z_0} * z_1$$
$$dy = y_1 - \frac{y_0}{z_0} * z_1$$

em que x_0, y_0 e z_0 correspondem às posições x, y e z, respectivamente, do sub-cluster do primeiro segmento e x_1, y_1 e z_1 correspondem às posições x, y e z, respectivamente, do sub-cluster do segundo segmento. Se dx e dy estão dentro do intervalo]-5,5[, calcula-se a distância entre o sub-cluster do segundo segmento e o do terceiro e assim sucessivamente. Estas equações são deduzidos da mesma maneira que as equações 5.2 e 5.3.

7 Após isso essas informações são salvas.

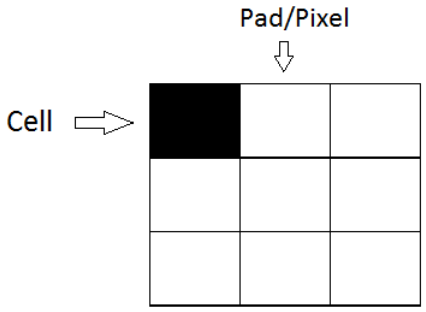


Figura 5.6: Representação transversal célula e pad.

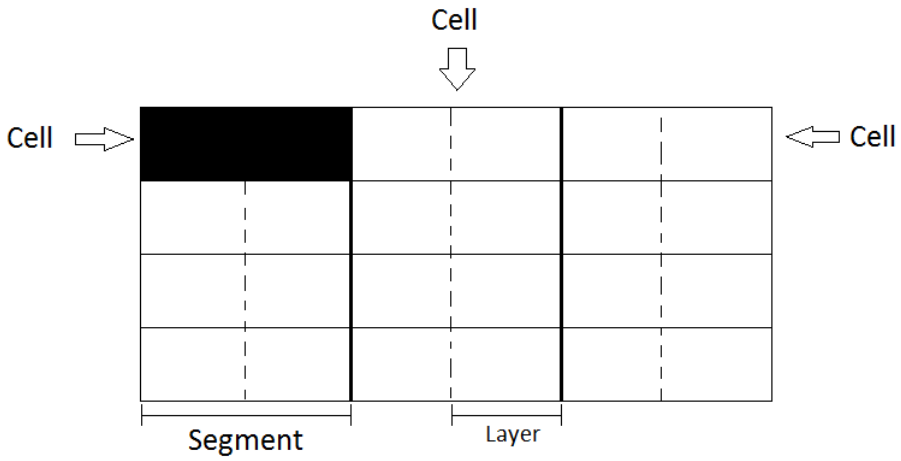


Figura 5.7: Representação da visão longitudinal da célula, segmento e layer.

Os resultados para o algoritmo com as modificações supramencionadas, estão apresentados na seção 6.2.

5.4.3 Algoritmo 3

Nesta etapa ocorreu a mudança para a Geometria 3 (apresentada resumidamente na seção 4.1) e com ela, uma outra modificação no algoritmo que foi alterar a maneira em que se cria o cluster dentro de um segmento. Inicialmente, escolhia-se uma matriz 3×3 ao redor da célula central. Agora, calcula-se uma distância entre a célula central e a célula que se quer unir. Primeiramente calcula-se para a diferença entre as posições delas para as direções x e y, obtendo assim:

$$\Delta X = X_{CELcentral} - X_{CEL} \tag{5.13}$$

$$\Delta Y = Y_{CELcentral} - Y_{CEL} \tag{5.14}$$

e após isso obtêm-se:

$$r = \sqrt{\Delta X^2 + \Delta Y^2} \tag{5.15}$$

sendo que $r < r_{min}$ a ser definido. Os resultados com esta nova maneira de se unir as células estão apresentados na seção 6.3.

Capítulo 6

Resultados e Discussão

Houve duas fases para as simulações, sendo que em ambas utilizou-se o calorímetro isolado (sem os outros detectores do ALICE). Na primeira fase foram realizadas simulações localmente, utilizando o FoCal Framework, para as energia de 1, 2, 6, 15, 34, 85, 200 GeV. A geometria utilizada foi a Geometria 1 (descritas na seção 6.1). Estas geometrias tiveram como base a primeira geometria apresentada na seção 4.1. Variou-se o ligeiramente o número de *layers* e o tamanho do *pixel*. Fazendo uma combinação entre estas obteve-se quatro geometrias. Nelas foi executado o algoritmo no estágio de desenvolvimento apresentado na seção 5.4.1. Assim, seriam obtidos resultados de como a geometria influenciaria na clusterização e quais poderiam ser suas modificações para que esta clusterização fosse eficiente. Estes resultados e a descrição das geometrias estão apresentados na seção 6.1. Após as modificações no algoritmo apresentadas na seção 5.4.2, testou-se este algoritmo na Geometria 2. Os resultados obtidos estão apresentados na seção 6.2.

Na segunda fase a colaboração decidiu padronizar as simulações para que os resultados obtidos por cada membro pudessem ser comparados de forma idônea. Esta simulação foi realizada da mesma maneira que a local, mas com a diferença de que ao invés de ser realizada para energias específicas, foi realizada para quatro valores de rapidez[1], sendo eles $y = 2.5$, 3.0, 3.5 e 4.0. Utilizou-se o algoritmo da seção 5.4.3 com a Geometria 3. Nesta parte também foram realizadas otimizações para este algoritmo afim de que seu desempenho fosse melhor para esta geometria. A descrição dela e os resultados obtidos estão apresentados na seção 6.3.

Na figura 6.1 é apresentada um modelo esquemático do calorímetro para facilitar a visualização das geometrias descritas nas seções abaixo.

[1]Para maiores detalhes sobre rapidez, vide apêndice B

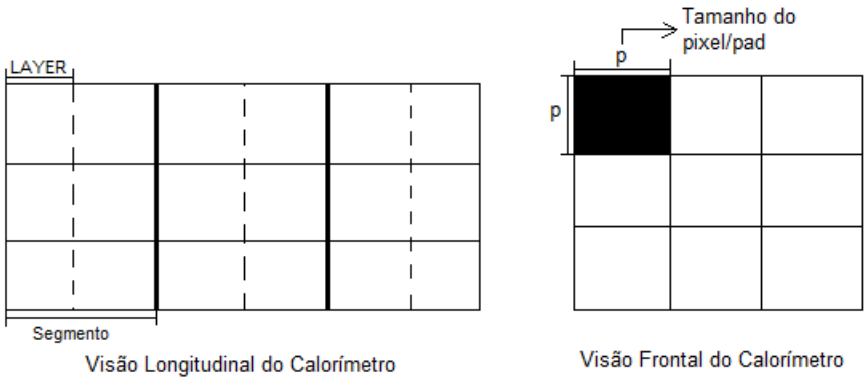


Figura 6.1: Modelo esquemático do calorímetro. Esquerda: Visão longitudinal. Direita: Visão frontal.

6.1 Geometria 1

Nesta parte, foi utilizado o algoritmo no estágio de desenvolvimento apresentado na seção 5.4.1. Foram utilizados dois tamanhos de pixel e dois números de *layers* que combinadas totalizam quatro geometrias diferentes. Observando a figura 6.1, abaixo estão apresentados os valores utilizados e na figura 6.2 estão esquematizados os detectores de 24 e 30 camadas.

- tamanho do pixel (p): 100 μm e 30 μm;
- número de layers: 24 e 30 *layers*;
- número de segmentos: zero.

Para todas as combinações destas 4 variáveis os materiais que compõe os *layer* são: uma camada de tungstênio (0,15 cm), um *pad* de silício (0,002 cm), cabos (0,02 cm), ar (0,078 cm) e uma última camada de tungstênio (0,15 cm). Esta disposição dos materiais no *layer*, está apresentado na figura 6.3.

Além disto foi variado o intervalo de dx e dy para relacionar os sub-clusters de difrentes segmentos. Este foi variado de 1 a 5 com passos de 1.

Para as figuras 6.4 a 6.6, os aglomerados de pontos azuis correspondem ao ponto em o que o fóton entrou no calorímetro, enquanto que os quadrados em pretos correspondem ao ponto que o algoritmo identificou como centro do cluster.

Podemos observar na figura 6.4 que para píons de 6 GeV num detector de 24 *layers*, não importando o tamanho do pixel, já não é possível separar os fótons provenientes do decaimento. Isto é devido ao tamanho do intervalo de busca para as varáveis dx e dy ser muito grande (]-5,5[). Para fótons (figuras 6.4(c) e 6.4(d)) é possivel ver que o algoritmo

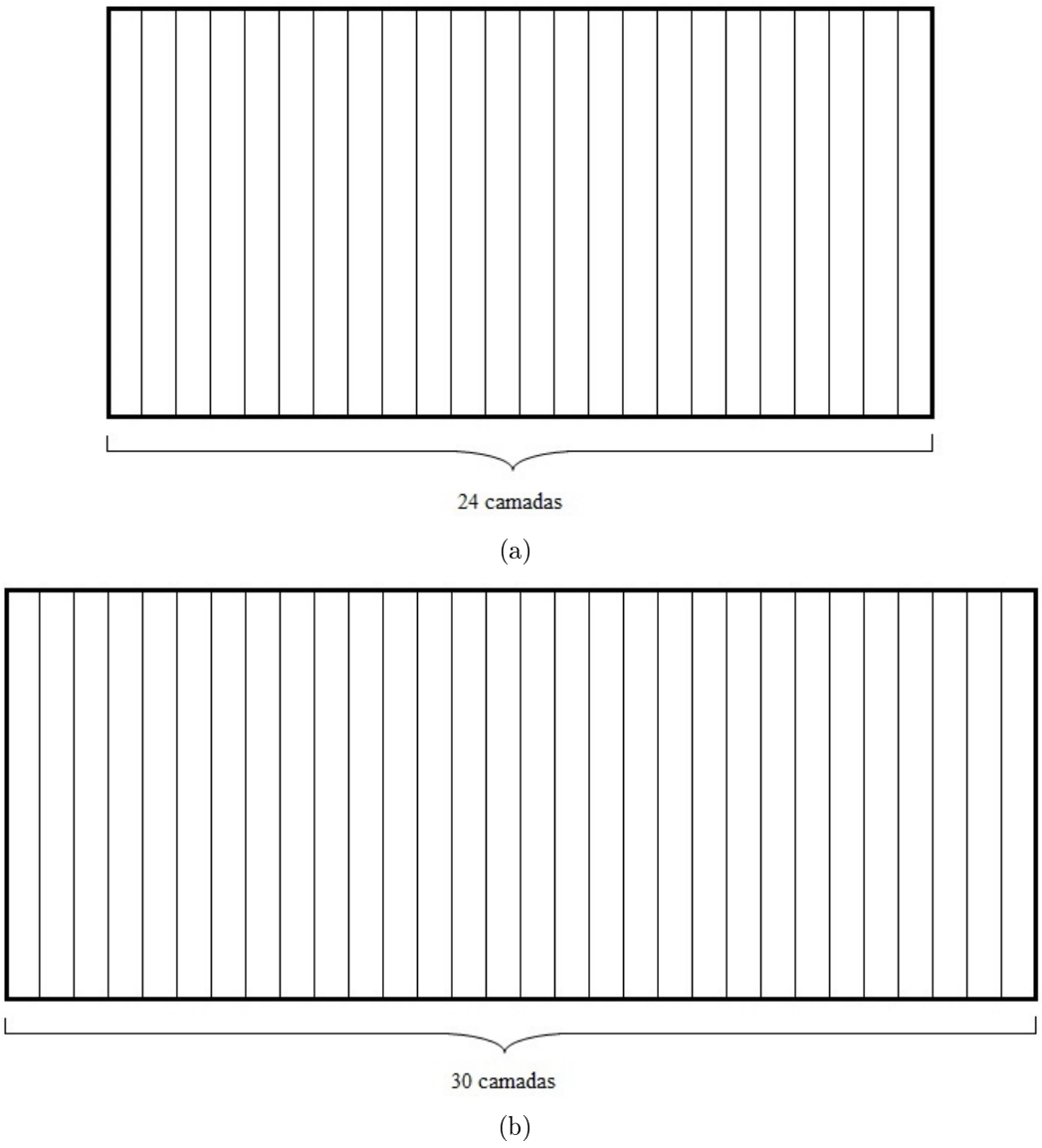


Figura 6.2: Esquema dos detectores. (a) Detector com 24 camadas (*layers*). (b) Detector com 30 camadas (*layers*).

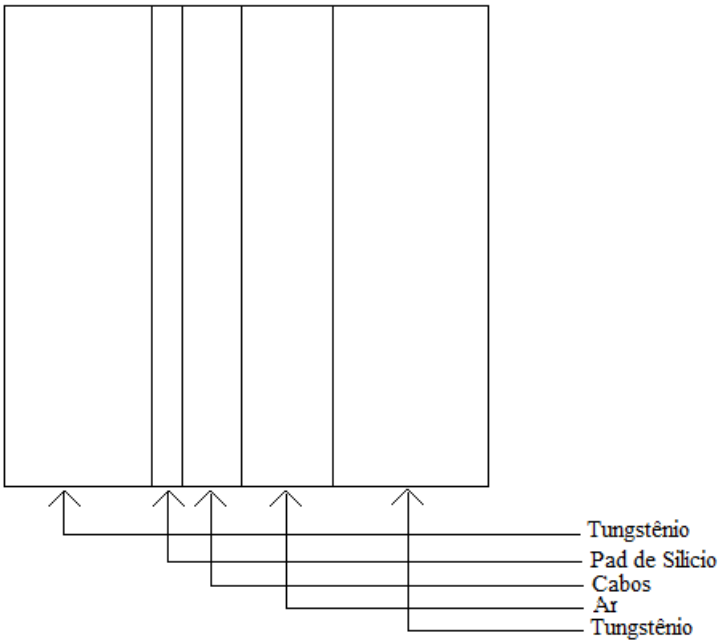


Figura 6.3: Visão longitudinal do *layer* da geometria 1.

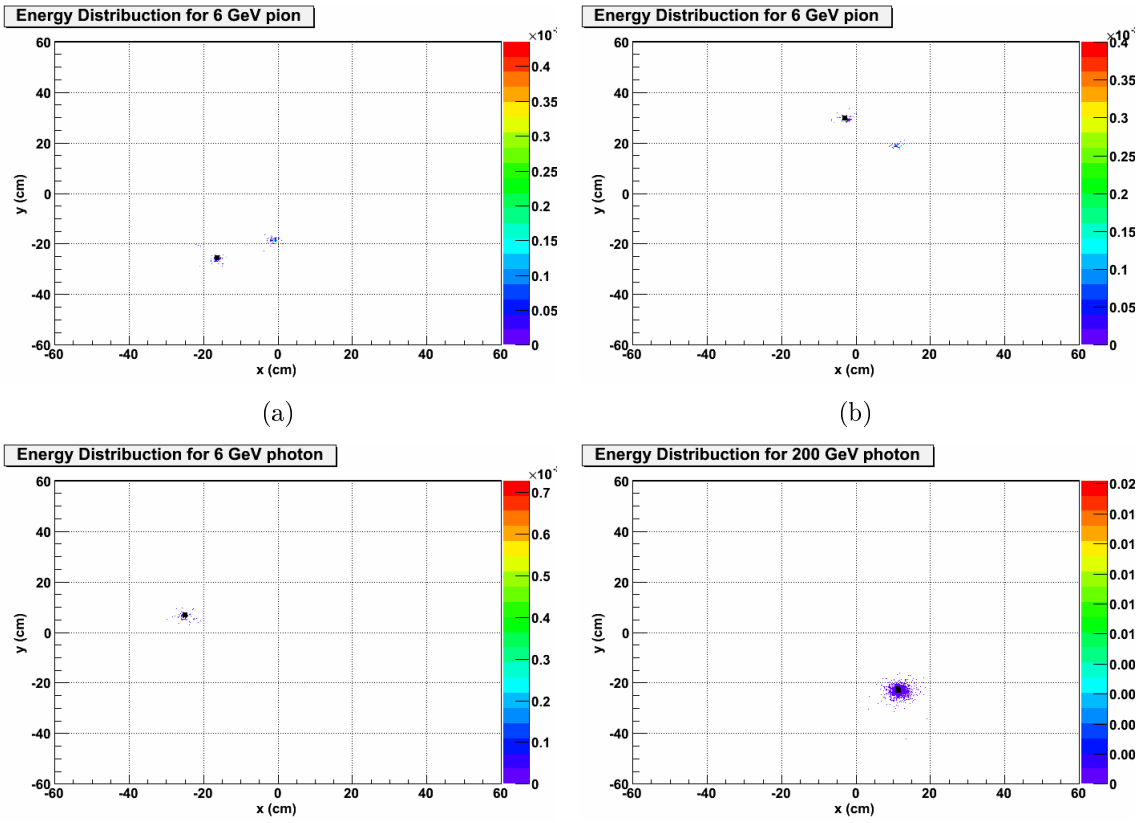
funciona razoavelmente bem para esta energia e às outras independente do tamanho do pixel.

Observando a figura 6.5, utilizando 30 *layers*, é possível notar a separação entre píons de 6 GeV. Esta separação vai até 15 GeV (para ambos os tamanhos de pixel), como mostrado na figura 6.5(c) e 6.5(d), o que não ocorre para a simulação com 24 *layers*. Para fótons o algoritmo também funciona bem (figura 6.5(e)).

Comparando as figura 6.4 e 6.5 pode-se notar que a melhora na detecção não depende do tamanho do pixel, mas sim do número de *layers*. Devido ao maior tamanho do detector maior a probabilidade da partícula depositar toda sua energia, e esta variável é fundamental para o algoritmo em uso. Pode-se observar na figura 6.6 que com a diminuição do intervalo para dx e dy ocorre um aumento na criação de artefatos. Como este intervalo é muito pequeno, o algoritmo não computa toda a energia depositada para um cluster, assim fica uma região na borda do chuveiro que pode dar origem a outros clusters.

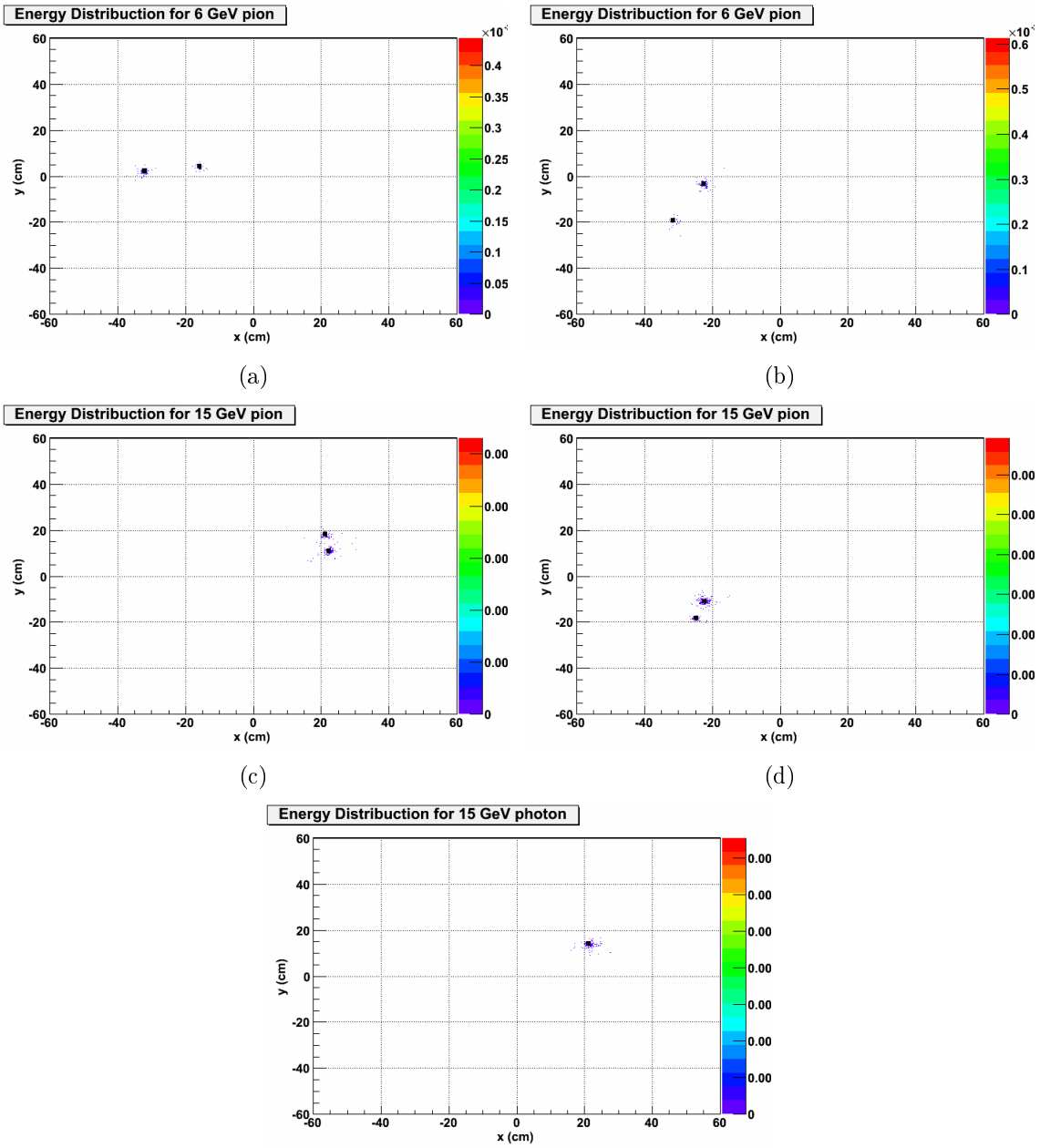
6.2 Geometria 2

Nesta parte, foi utilizado o algoritmo no estágio de desenvolvimento apresentado na seção 5.4.2. Após os testes iniciais com a geometria da seção anterior, iniciou-se um pequeno estudo de como este algoritmo se comportaria para geometria desenvolvida por Taku Gunji. Observando, novamente, a figura 6.1, a geometria é:



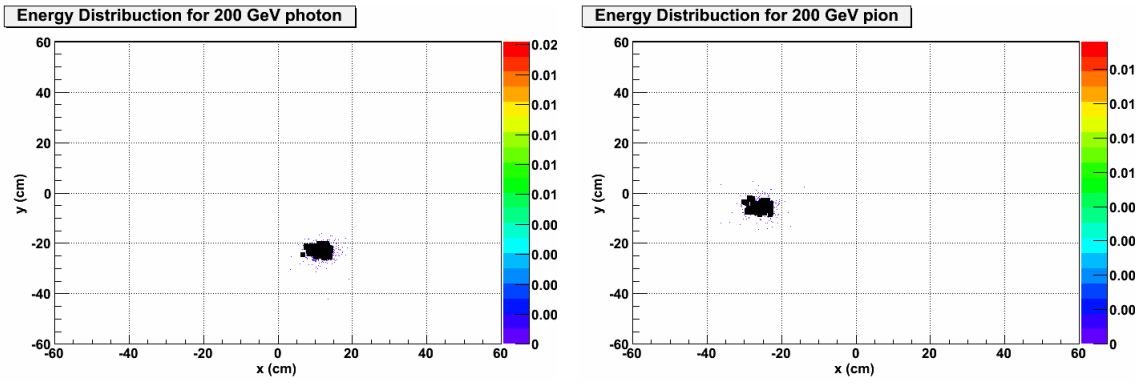
(a) (b) (c) (d)

Figura 6.4: Resultados para detector com 24 *layers* com dx e dy dentro do intervalo de [-5,5]. Os quadrados pretos são o centro do cluster encontrado pelo algoritmo. (a) Píons de 6 GeV com pixel do tamanho de 30 μm. (b) Píons de 6 GeV com pixel do tamanho de 100 μm. (c) Fóton de 6 GeV com pixel do tamanho de 30 μm. (d) Fóton de 200 GeV com pixel do tamanho de 100 μm.



(a) (b) (c) (d) (e)

Figura 6.5: Resultados para detector com 30 *layers* com dx e dy dentro do intervalo de [-5,5]. Os quadrados pretos são o centro do cluster encontrado pelo algoritmo. (a) Píons de 6 GeV com pixel do tamanho de 30 μm. (b) Píons de 6 GeV com pixel do tamanho de 100 μm. (c) Píon de 15 GeV com pixel do tamanho de 30 μm. (d) Píon de 15 GeV com pixel do tamanho de 100 μm. (e) Fóton de 15 GeV com pixel do tamanho de 30 μm.



(a) (b)

Figura 6.6: Resultados para detector com 30 *layers* com dx e dy dentro do intervalo de [-1,1]. Os quadrados pretos são o centro do cluster encontrado pelo algoritmo. (a) Fóton de 200 GeV com pixel do tamanho de 100 μm. (b) Píons de 200 GeV com pixel do tamanho de 100 μm.

- tamanho do pixel (p): 0,0055cm;
- número de layers: 21 *layers*;
- número de segmentos: 3 com 7 *layers* cada.

Na figura 6.7 está esquematizado o detector acima.

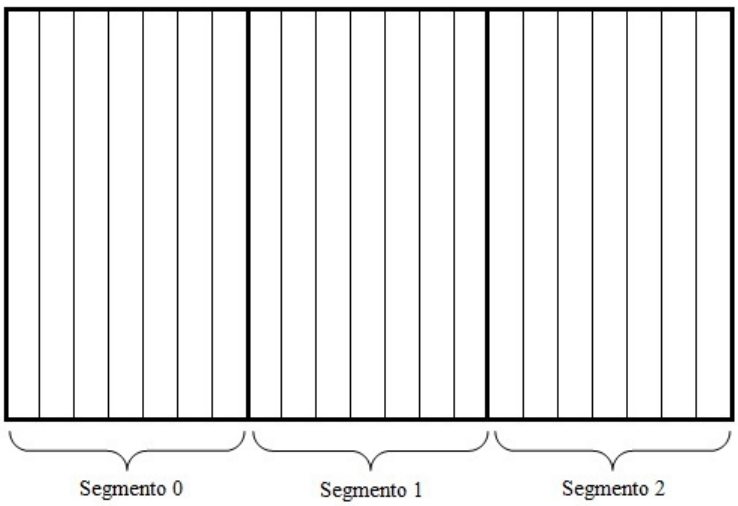


Figura 6.7: Esquema do detector. Ele tem 21 camadas (*layers*) e é dividido em três segmentos.

Cada *layer* tem a seguinte composição: *Alloy* (mistura de vários dois ou mais elementos - 0.35cm), G10 (0,01 cm), *pad* de silício (0,0535 cm), G10 (0,01), ceramica (0,08 cm) e ar (0,1 cm). A simulação deste detector foi realizada para 100 eventos, tanto píons quanto

fótons. A disposição dos componentes deste *layer* está na figura 6.8. Para esta geometria também foi variado o intervalo de dx e dy para relacionar os sub-clusters de difrentes segmentos (tópico 6 da seção 5.4). Este foi variado de 1 a 5 com passos de 1.

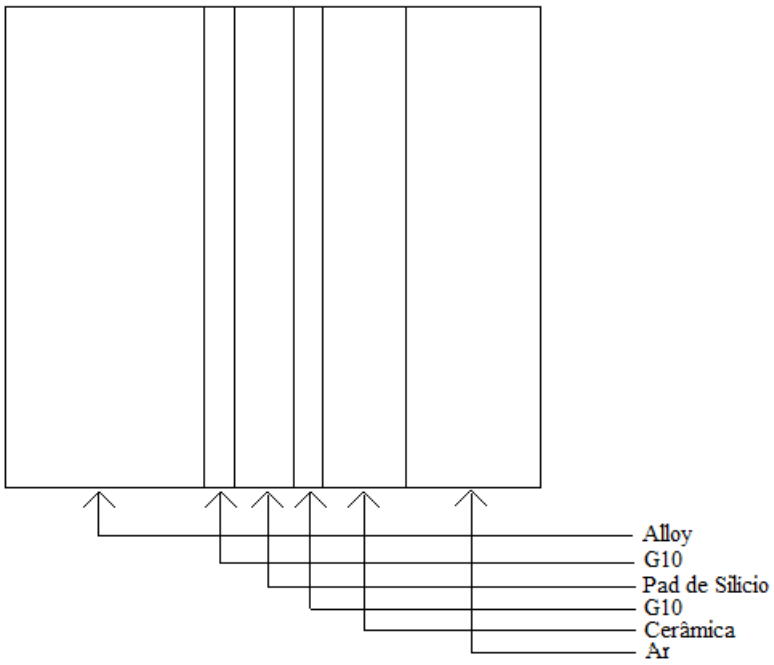


Figura 6.8: Visão longitudinal do *layer* da geometria 2.

Abandonando a análise qualitativa da seção anterior, obtemos na figura 6.9, histogramas para o número de *clusters* encontrados num conjunto de 100 eventos para fótons. Nestes gráficos pode-se observar que para energias mais baixas o algoritmo encontra somente um *cluster* (figura 6.9(a)) , o que é o esperado. Conforme aumenta-se a energia começa a aparecer um segundo pico no ponto de 2 *clusters* (figuras 6.9(b) e 6.9(c)). Para a energia de 250 GeV (figura 6.9(d)) além das contagens em 2 *clusters* estarem altas, há também uma contagem considerável para 3 *clusters*. Isto tende a ocorrer conforme aumenta-se a energia da partícula, pois a largura de seu chuveiro aumenta e com isso aumenta o número de células com maior energia encontradas por segmento.

Observando a figura 6.10, para píons, pode-se notar o seguinte comportamento: para energias mais baixas (figura 6.10(a)) o algoritmo consegue encontrar dois *clusters*. Conforme aumenta-se a energia(figuras 6.10(b) e 6.10(c)) nota-se um comportamento oposto ao do fóton. Ao invés de se aumentar o número de *clusters* encontrado, tem-se uma diminuição do mesmo. Isto ocorre pois conforme aumenta-se a energia a separação dos fótons emitidos pelo píon se torna menor a ponto do algoritmo não conseguir fazer distinção entre os dois e considerá-lo um único *cluster*. Para 250 GeV (figura 6.10(d)) começa a aparecer uma contagem maior em dois *clusters*, mas isto é devido ao chuveiro ser largo,

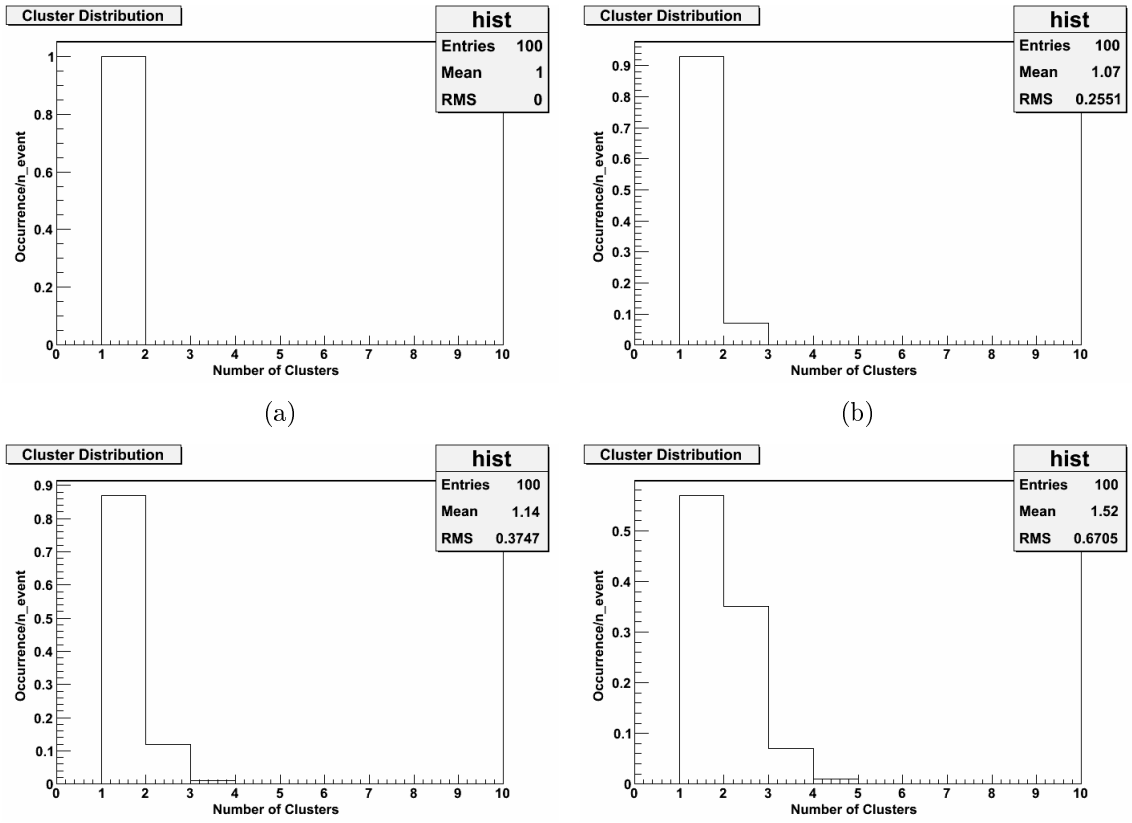


Figura 6.9: Histrogramas para o número de clusters encontrado para cada evento para fótons (dx e dy dentro do intervalo de [-5,5]). (a) Energia de 10 GeV. (b) Energia de 50 GeV. (c) Energia de 100 GeV. (d) Energia de 250 GeV.

dada a energia do píon, e não à capacidade do algoritmo em discriminar os dois fótons.

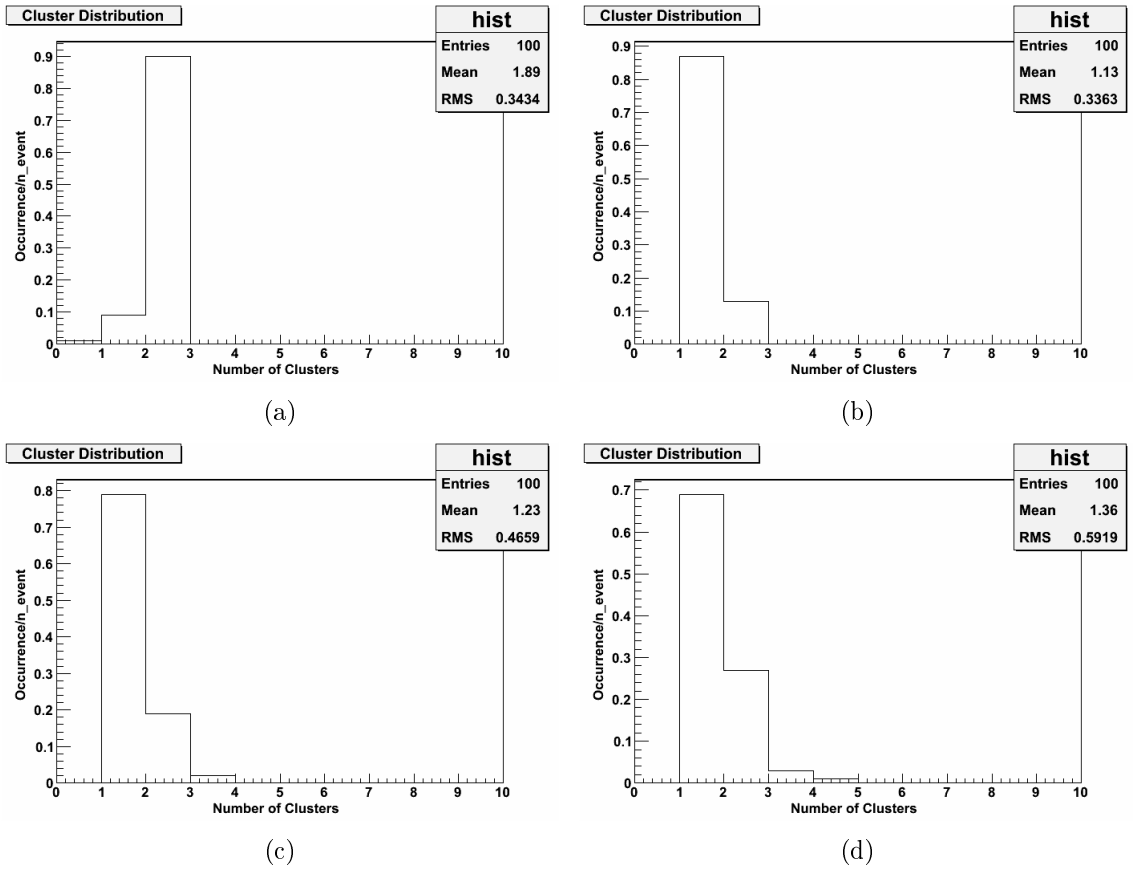
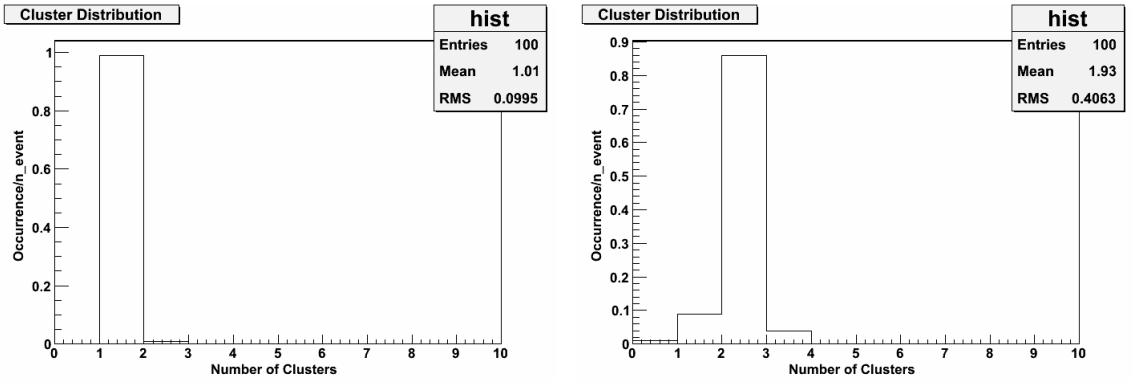


Figura 6.10: Histrogramas para o número de clusters encontrado para cada evento para píons (dx e dy dentro do intervalo de [-5,5]). (a) Energia de 10 GeV. (b) Energia de 50 GeV. (c) Energia de 100 GeV. (d) Energia de 250 GeV.

De acordo com a figura 6.11 é possível perceber que a variação do intervalo de dx e dy pouco influencia no número de *clusters* encontrado. Com isso pode-se concluir que esta variável do algoritmo não tem um papel fundamental na busca pelos *clusters* corretos.

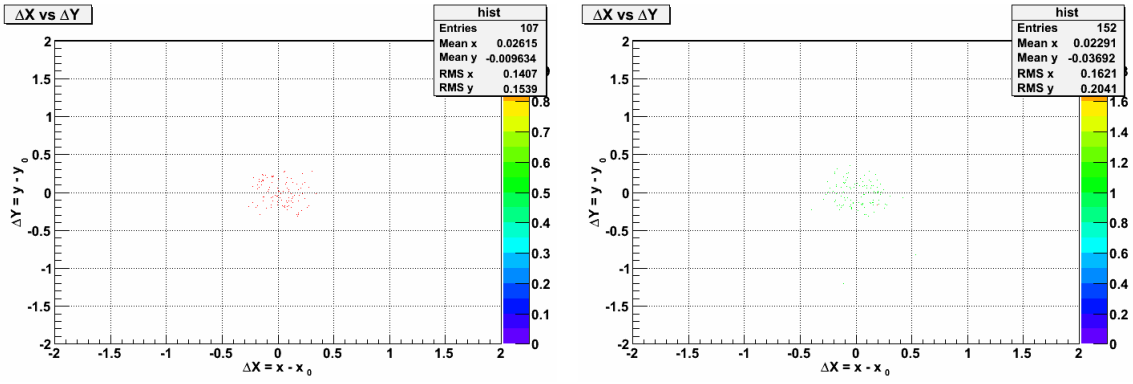
Agora, olhando para a distância em Y e em X do ponto onde foi encontrado o *cluster* (figura 6.12), para o ponto em que o fóton incidiu no detector, é possível perceber que há uma variação de quase um centímetro em cada eixo. Isto mostra que apesar do algoritmo encontrar o *cluster* ele não está na posição correta. Olhando para outras energias podemos concluir que esta dispersão independe da energia da partícula incidente.

Colocando num gráfico os valores médios dos histogramas das figuras 6.9 e 6.10 e seus respectivos erros obtemos a figura 6.13. Nela podemos observar que, para fótons, temos um aumento no número de clusters com o aumento da energia. Para píons, vemos que baixas energias o algoritmo tende a encontrar dois *clusters*, mas ao aumentar a energia este número cai rapidamente para valores próximos de um. A energias muito baixas o algoritmo encontra um único cluster pois os fótons do decaimento tem seu ângulo muito



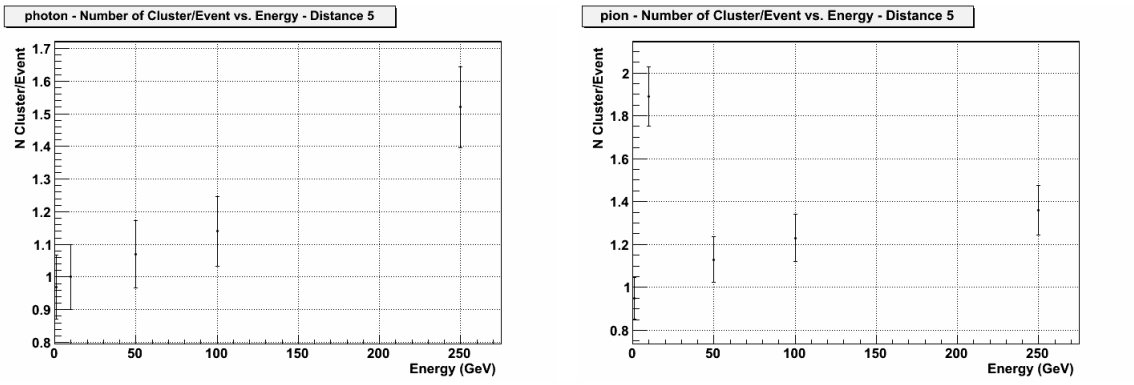
(a) (b)

Figura 6.11: Histrogramas para o número de clusters encontrado para cada evento para energia de 10 GeV (dx e dy dentro do intervalo de [-1,1]). (a) Fóton. (b) Píon.



(a) (b)

Figura 6.12: Figura para $\Delta X = X - X_0$ (cm) e $\Delta Y = Y - Y_0$ (cm), em que (X, Y) é o ponto de entrada do fóton e (X_0, Y_0) é o ponto do centro do *cluster*. (a) Energia de 50 GeV. (b) Energia de 250 GeV.



(a) (b)

Figura 6.13: Figura para número de *cluster* pelo número de eventos *vs.* energia. (a) Fóton. (b) Píon.

grande levando a um dos deles não atingir o detector.

6.3 Geometria 3

Nesta parte, foi utilizado o algoritmo no estágio de desenvolvimento apresentado na seção 5.4.3. Utilizando como modelo a figura 6.1 para a geometria mais atual do FoCal (seção 4.1), tem-se que:

- tamanho do pixel/pad (p): $1cm$ (LGL) e $0{,}01cm$ (HGL) combinadas em pad de $0.1cm$ (macropixel);
- número de segmentos: 6. O segmento 0 é composto por 4 *layers* LGL. Este segmento é igual ao segmento 2. O segmento 1 é composto por um único *layer* HGL. Este segmento é igual ao segmento 3. Os segmentos 4 e 5 são iguais, sendo formado por 5 *layers* LGL cada.

Na figura 6.14 está esquematizado o detector acima.

Cada *layer* é composto por tungstênio (0.35 cm), G10 (0.05 cm), *pad* de silício (0.05 cm), novamente o G10, cobre (0.01 cm) e ar (0.05 cm). A colocação destes materiais no *layer* está representada na figura 6.15.

Seguindo o procedimento do tópico anterior, na figura 6.16 obtem-se o número de clusters encontrados por evento. Como pode-se observar, para esta geometria, o algoritmo encontra uma quantidade grande de clusters que não correspondem ao fóton que produziu o chuveiro.

Devido a este problema, pensou-se em algumas modificações. Uma delas foi, ao invés de se considerar uma matriz 3×3, passou-se a calcular o raio ($r = \sqrt{\Delta X^2 + \Delta Y^2}$) e este

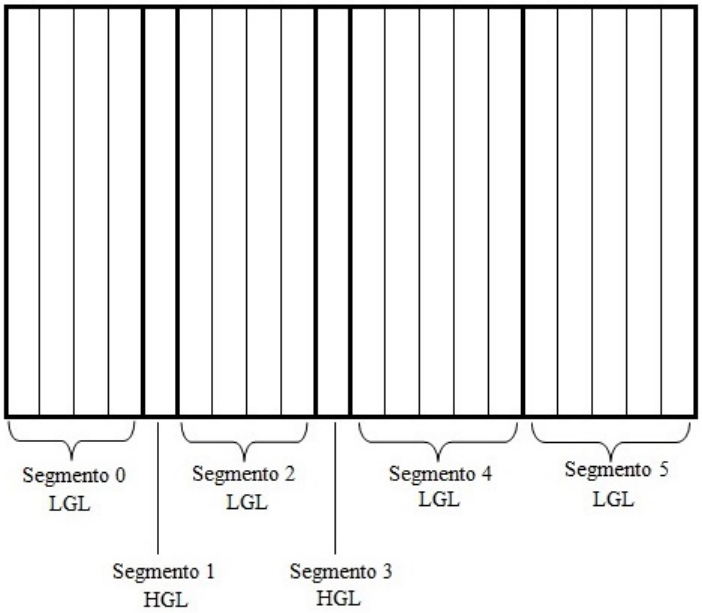


Figura 6.14: Esquema do detector apresentando as disviões dos segmentos e a disposição das camadas (*layers*) de HGL e LGL.

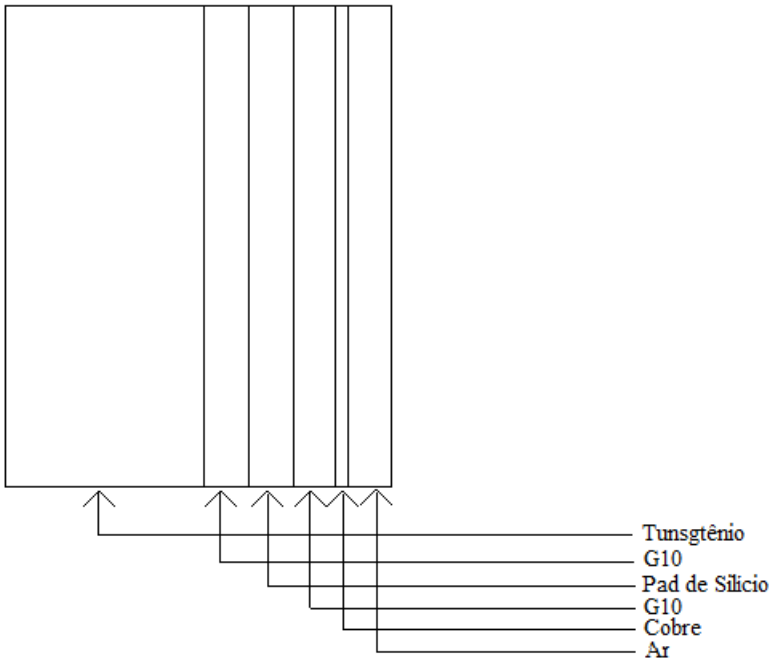
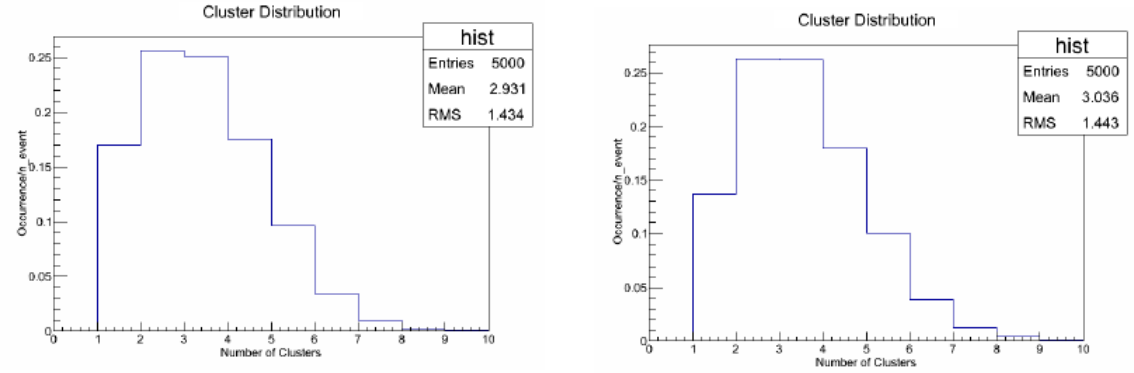


Figura 6.15: Visão longitudinal do *layer* da geometria 3.



(a) (b)

Figura 6.16: Número de clusters por evento para rapidez $y = 3.0$. (a) Fóton. (b) Píon.

deve ser o menor entre os sub-clusters de cada segmento. Outro ponto, foi procurar a otimização de três parâmetros utilizados no algoritmo. Elas são:

- Energia mínima para considerar uma célula na criação do *cluster*;
- Energia mínima para considerar uma célula como centro do *cluster*;
- Distância mínima para procurar por células vizinhas.

6.3.1 Energia mínima para considerar uma célula na criação do *cluster*

Para esta variável calculou-se a energia depositada em cada célula, obtendo-se a figura 6.17. Nela podemos observar que para energias muito baixas há um pico. Este está relacionado a regiões da borda do chuveiro onde a energia depositada ali é muito baixa. Após este pico temos um vale seguido de outro pico, como mostrado, por exemplo, nas figuras 6.17(a) e 6.17(b). O corte foi tomado na região próxima à subida do segundo pico. No caso do seguimento 1 (seguimento de HGL) só há um único pico, o corte então foi tomado na região, também próxima à subida deste pico. Nas figuras 6.17(c) e 6.17(d) também observa-se que o valor do corte utilizado para rapidez diferentes é o mesmo.

6.3.2 Energia mínima para considerar uma célula como centro do *cluster*

Nesta parte fez-se um gráfico da energia de cada célula que o algoritmo denominou como máximo (executou-se esta parte com ausência de corte) pela distância euclidiana entre ponto onde o fóton entrou no detector e o ponto onde esta se localizava ($r =$

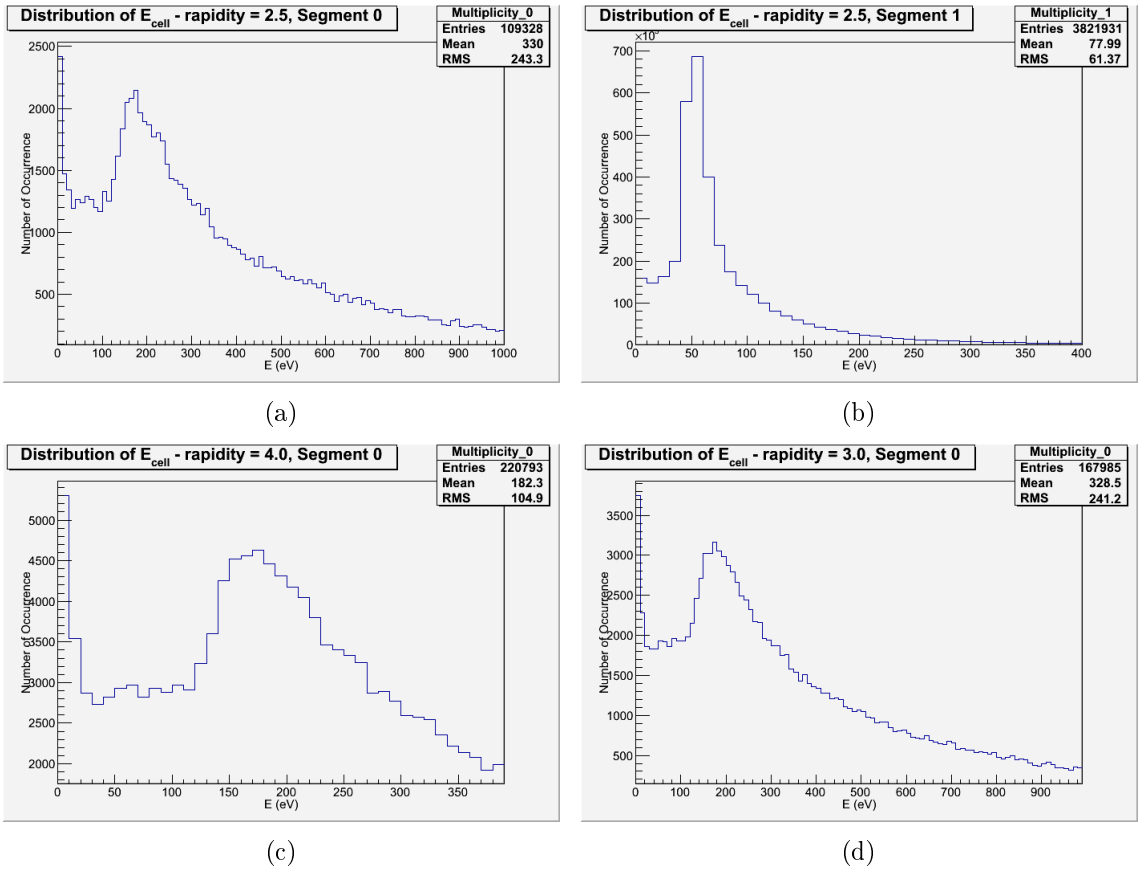


Figura 6.17: Energia depositada em cada célula. (a) Fóton, segmento 0 para y = 2.5. (b) Fóton, segmento 1 para y = 2.5. (c) Fóton, segmento 0 para y = 4. (d) Píon, segmento 0 para y = 3.0.

$\sqrt{(X_{cell} - X_\gamma)^2 + (Y_{cell} - Y_\gamma)^2}$). Com isto obteve-se a figura 6.18. Nela observa-se que a deposição da energia depende do tamanho do pad/pixel (os segmentos 1 e 2 têm pads de tamanho diferente). É possível notar uma região com pontos bem próximo de zero, uma região com praticamente nenhuma contagem seguida de outra região com vário pontos. O objetivo era eliminar o máximo possível os pontos da região mais distante de zero. Este corte foi escolhido de uma maneira que não fosse muito restritivo (começaria a perder células centrais que corresponderiam aos fótons que entraram no detector) e nem muito brando (aceitando assim, muitas células que não correspondem a nada real), assim nas figuras 6.18(c) e 6.18(d) mostra-se que o corte escolhido para um valor de rapidez é o mesmo para outros.

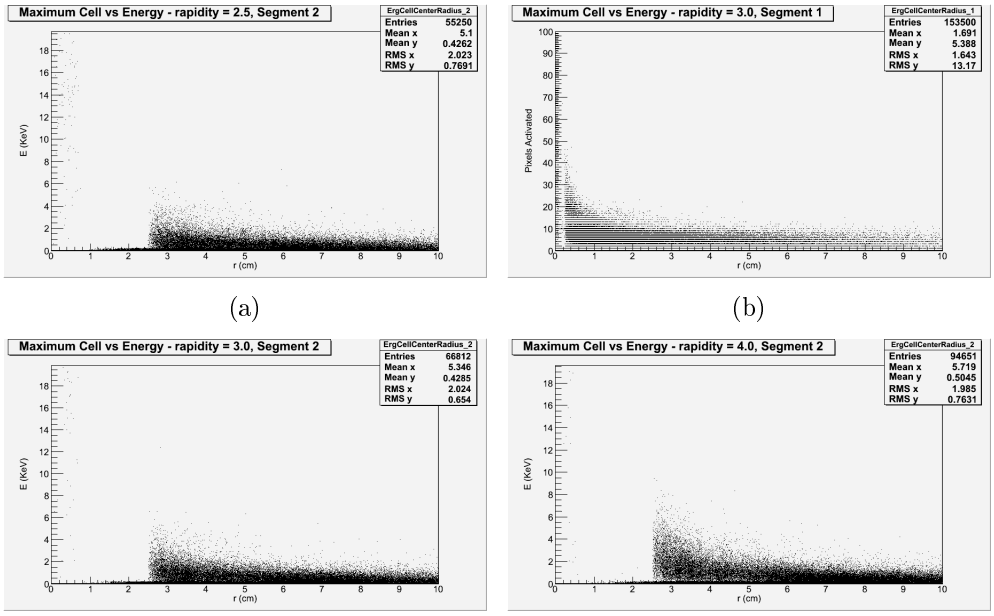


Figura 6.18: Energia depositada em cada célula tida como máximo *vs.* r. (a) Fóton, segmento 2 para y = 2.5. (b) Fóton, segmento 1 para y = 3.0. (c) Fóton, segmento 2 para y = 3.0. (d) Fóton, segmento 2 para y = 4.0.

6.3.3 Distância mínima para procurar por células vizinhas

Para esta parte calculou-se a distância entre os fótons provenientes do píon. Para isto realizou-se o seguinte procedimento: assumindo-se que o fóton a partir de sua criação anda numa linha reta, calculou-se sua posição no lugar em que ele acertaria cada segmento do detector. Assim, observando a figura 6.19 temos o momento de um dos fóton no plano YZ na origem (Ponto de Interação - PI). O ângulo formado por este e o eixo Z é α_1.

Têm-se que:

$$tg\alpha_1 = \frac{P_Y}{P_Z} = \frac{Y_1}{Z} \Rightarrow Y_1 = Z\frac{P_Y}{P_Z} \Rightarrow Y_1 = Ztg\alpha_1 \tag{6.1}$$

Analogamente para o plano XZ têm-se:

$$tg\theta_1 = \frac{P_X}{P_Z} = \frac{X_1}{Z} \Rightarrow X_1 = Z\frac{P_X}{P_Z} \Rightarrow X_1 = Ztg\theta_1 \tag{6.2}$$

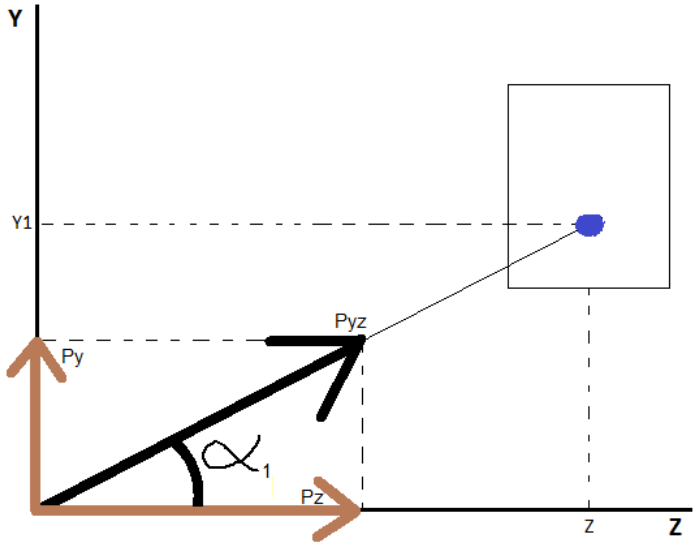


Figura 6.19: Esquema com a disposição das variáveis de interesse.

Para a segunda partícula o procedimento é o mesmo, mas trocando o 1 por 2. Assim, observando a figura 6.20, deve-se calcular a distância D entre os fótons do decaimento, que é dada por,

$$D = \Delta S_\gamma = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2} = \sqrt{(Ztg\alpha_1 - Ztg\alpha_2)^2 + (Ztg\theta_1 - Ztg\theta_2)^2} \tag{6.3}$$

Na figura 6.21 pode-se ver o resultado deste cálculo em função da energia inicial do píon. Nela observa-se que o comportamento e os valores são semelhantes para outros valores de rapidez.

Nela pode-se observar que para os diferentes segmentos, a separação dos fótons não muda muito. Com isso foram escolhidos os seguintes valores para os raios de busca: 3 cm para os segmentos 0, 2, 4 e 5 (LGL) e 0.3 cm para os segmentos 1 e 3 (HGL). Pode-se argumentar observando, principalmente, a figura 6.21(c) que se a distância de separação para 1000 GeV é de 0.1 cm por que foi escolhido 0.3 cm? Ao se recordar que a menor unidade sensível deste segmento (segmento 1) é de $0.1cm$, se fosse escolhido este tamanho, cada macropixel seria um cluster individual, o que tornaria o algoritmo inútil pois haveria

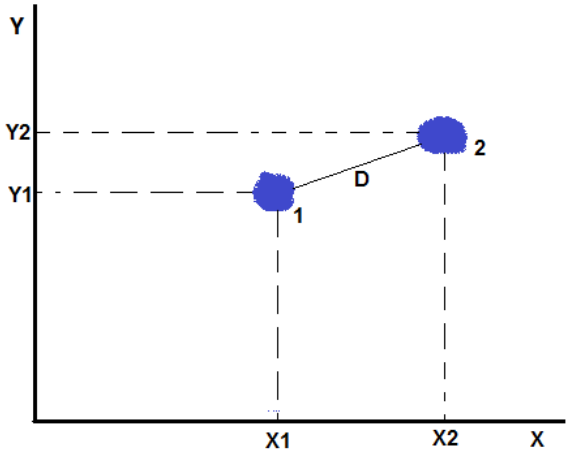


Figura 6.20: Visão frontal dos fótons de decaimento.

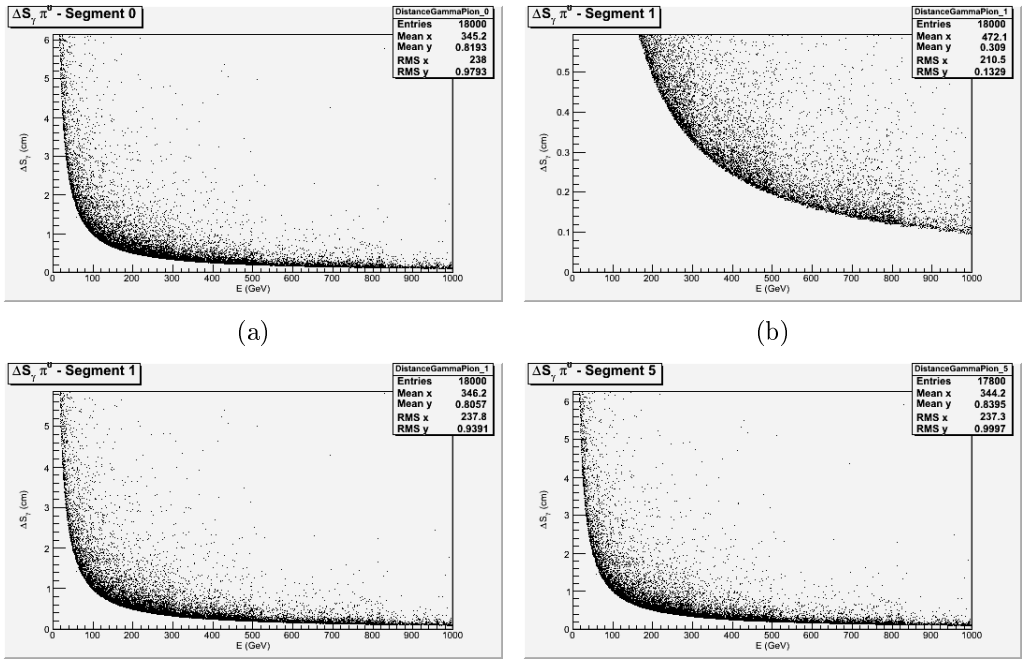


Figura 6.21: Separação dos fótons *vs.* energia do píon. (a) Segmento 0. (b) Segmento 1 com enfase na parte onde se escolheu o limite. (c) Segmento 1 mostrando o mesmo comportamento dos outros segmentos. (d) Segmento 5.

centenas de clusters sem nenhuma relação com o fóton real. Além do mais, com este raio, é possível somar ao cluster mais energia, tendo assim uma energia mais próxima da realmente depositada pelo o fóton. Analisando os segmentos LGL a menor unidade sensível é de 1 cm, então os clusters com separação menor que esta não serão identificados uma vez que pode-se utilizar os mesmos argumentos para os segmentos HGL para se

explicar a razão da escolha do valor de $3cm$.

Fazendo um pequeno exercício, no caso em que este valor de raio fosse menor, ter-se-ía a separação de clusters com energias superiores. Isto não seria um benefício sem custos, pois quando houvesse chuveiros mais largos, os clusters encontrados não cobririam toda a área deste chuveiro, o que isto levaria a criação de muitos clusters falsos em suas bordas com, possivelmente, energias comparáveis a dos clusters reais. No caso de um raio maior, um problema considerável seria uma menor capacidade de separação de fótons de decaimento de energias mais altas.

6.3.4 Resultados das Alterações

Com resultados obtidos nas seções anteriores têm-se os valores otimizados para os três parâmetros apresestados no início desta seção. Eles estão exibidos na tabela 6.1.

Tabela 6.1: Valores ajustados dos parâmetros.

Segmentos	**Energia Mínima da Célula Central (eV)**	**Energia Mínima da Célula (eV)**	**Raio de Busca (cm)**
0	3000	100	3,0
1	500	30	0,3
2	500	30	3,0
3	6000	100	0,3
4	50000	100	3,0
5	25000	100	3,0

Com isso, foram refeitas as análises obtendo-se a figura 6.22. Nesta figura é possível notar comportamento similar ao das figuras 6.9 e 6.10 da geometria 2. Para fótons pode-se observar que conforme aumenta-se a rapidez aumenta o numero de *clusters* falsos criados. Para píons, o algoritmo não é capaz de encontrar corretamente os *clusters* dos fótons de decaimento.

No gráfico de dispersão (figura 6.23), para fótons, é possível observar que a maior parte dos *clusters* encontrados estão bem próximos do ponto de entrada do fóton. A presença do padrão em cruz é, possivelmente, originária de *clusters* falsos (estes estão longe do ponto de entrada do fóton, pois se localizam mais na borda do chuveiro). Para píons, é possível notar que o algoritmo ainda tem dificuldade de encontrar os *clusters* dos fótons de decaimento de π^0, como mostrado em seus gráficos de dispersão, figuras 6.23(b) e 6.23(d), que são mais largos que os das figuras 6.23(a) e 6.23(c).

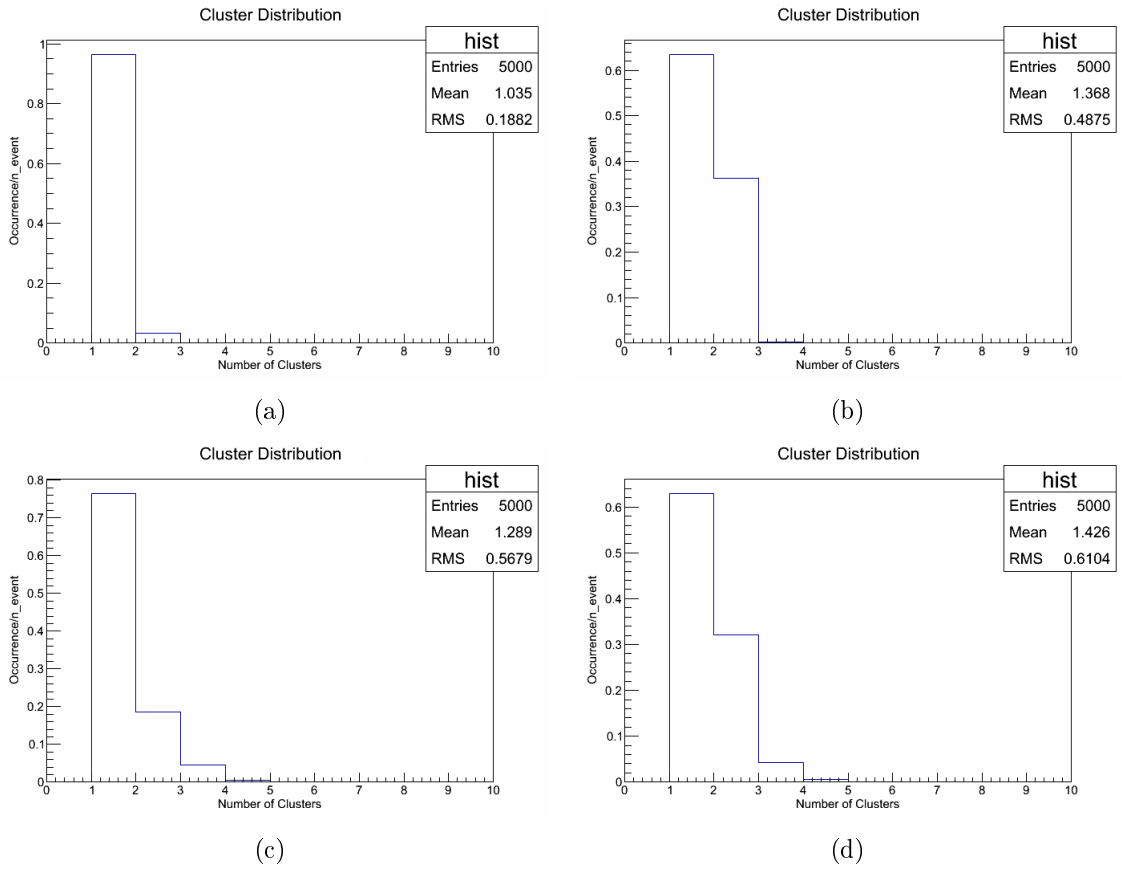


Figura 6.22: Histrogramas para o número de clusters encontrado para cada evento. (a) Fóton, y = 2.5. (b) Píon, y = 2.5. (c) Fóton, y = 3.5. (d) Píon, y = 3.5.

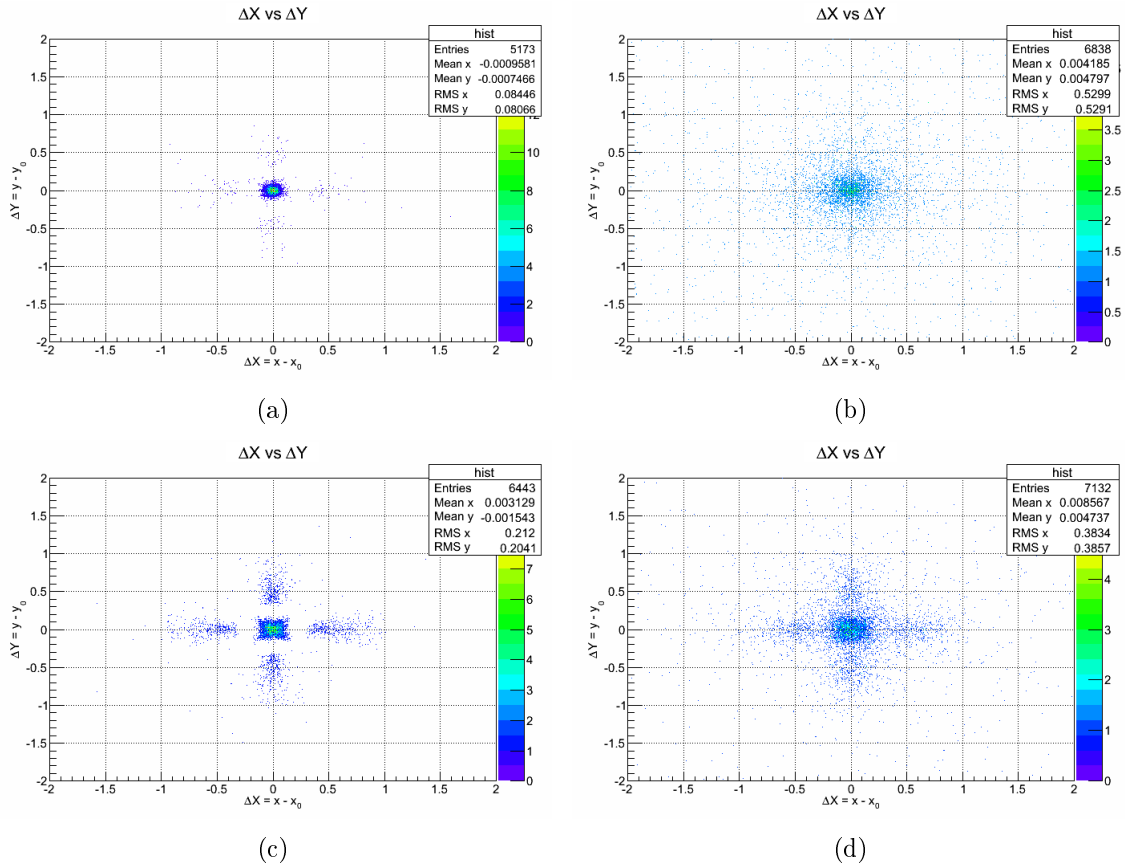


Figura 6.23: Figura para $\Delta X = X - X_0$ (cm) e $\Delta Y = Y - Y_0$ (cm), em que (X, Y) é o ponto de entrada do fóton e (X_0, Y_0) é o ponto do centro do *cluster*. (a) Fóton, y = 2.5. (b) Píon, y = 2.5. (c) Fóton, y = 3.5. (d) Píon, y = 3.5.

6.4 Cálculo da eficiência

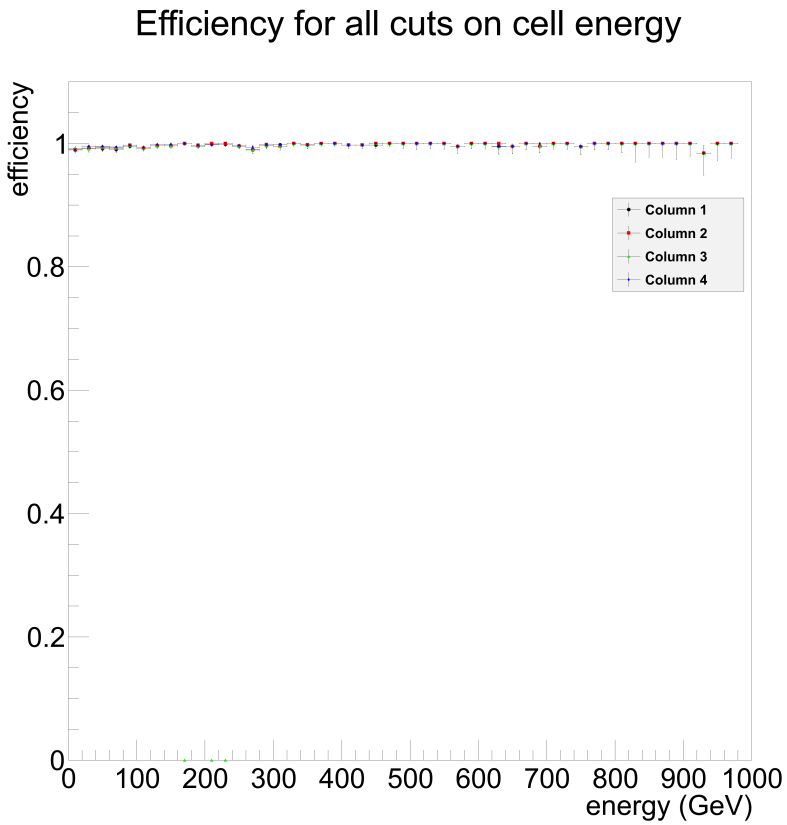
Com a otimização obtida na seção anterior, pensou-se em variar estes valores, calcular e comparar as eficiências na obtenção de clusters reais (clusters que correspondem ao fóton que atingiu o calorímetro). pelo algoritmo. Isto foi realizado da seguinte maneira: primeiramente criou-se um histograma com a energia inicial de todos os fótons (píons) produzidos na simulação; posteriormente realizou-se uma varredura, em todos os clusters encontrados em um evento, para a procura do cluster mais energético. Este método foi escolhido, pois assumiu-se que haveria um cluster bem energético que representaria a partícula incidente, e que os outros seriam falsos, pois teriam surgido de pequenas flutuações na borda do chuveiro (região que se encontra fora da distância mínima para se gerar um cluster). Para os píons foi utilizado o mesmo raciocínio, mas em vez de um cluster foi-se a procura dos dois clusters mais energéticos. Sendo assim, obteve-se os gráficos 6.24(a), para eficiência na identificação de fótons, e 6.24(b), para eficiência na identificação de píons. Estes gráficos são para os valores de cortes apresentados na tabela 6.2 e utilizando o cluster resultante da combinação dos sub-clusters encontrados em cada segmento. Como pode se verificar, a eficiência para píons está muito abaixo da de fótons.

Tabela 6.2: Valores dos cortes utilizados. Todos os cortes estão apresentados em eV, somente os segmentos 2 e 4 são dados em número de pixels ativados.

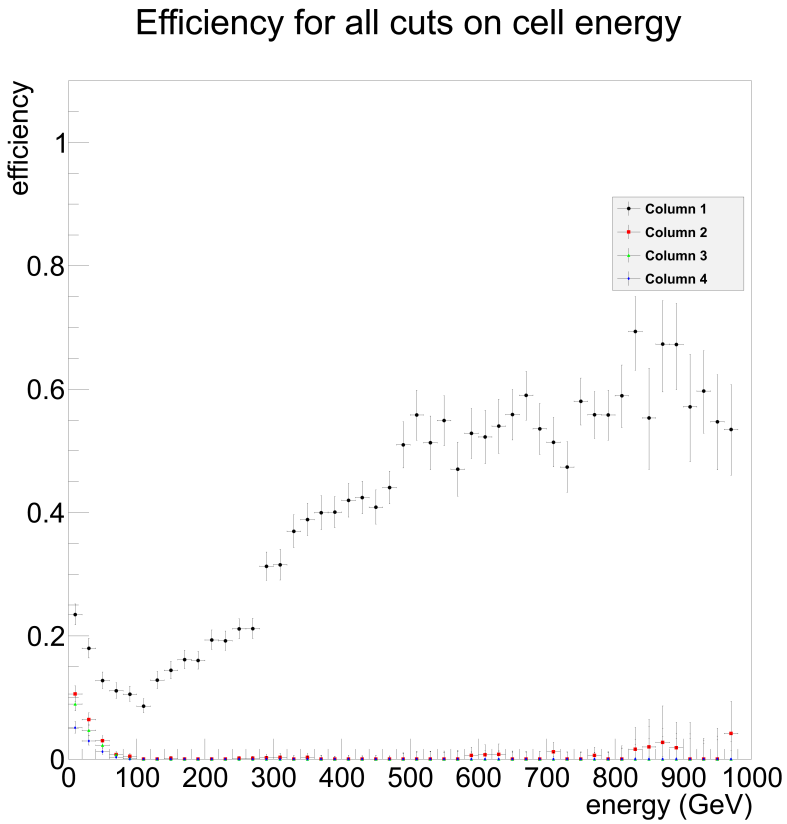
Segmentos	**Energia Mínima da Célula Central**			
	Corte 1	**Corte 2**	**Corte 3**	**Corte 4**
0	1000	2000	4000	10000
1	8	16	22	30
2	1500	4000	8000	12000
3	8	16	22	30
4	2000	8000	16000	22000
5	2000	8000	16000	22000

Como a eficiência para se identificar π^0 foi muito baixa, pensou-se em outras modificações para o algoritmo. Em vez de procurar unir os clusters do primeiro segmento ao último, realizou-se o seguinte: utilizou-se como posição dos clusters finais as obtidas pelos sub-clusters nos segmentos HGL. Os segmentos de LGL foram utilizados para se obter a energia depositada pelo chuveiro. Assim, deu-se início ao segundo cálculo de eficiência do algoritmo. Para isso foi utilizado o cálculo de massa invariante, pois com este teria-se uma maneira de identificar os clusters de fótons que decaíram de π^0.

Inicialmente utilizou-se os resultados da clusterização de fótons para se fazer calibração da energia. Esta foi executada da seguinte maneira: fez-se a relação entre a energia do cluster encontrado (energia depositada pelo chuveiro) e a energia inicial da partícula incidente. Com isso obteve-se quatro gráficos,apresentados na figura 6.25, um para cada



(a)



(b)

Figura 6.24: Eficências calculadas, utilizando a combinação de todos os segmentos, para os valores da tabela 6.2. (a) Eficência para identificação de γ. (b) Eficência para identificação de π^0.

corte da tabela 6.2 . O comportamento da figura 6.25(a) pode ser explicado pelo fato de que, como o corte foi muito brando, houve a inclusão de células com energia muito baixa (as quais estariam nas bordas do chuveiro) o que levou a uma geração excessiva de clusters falsos com baixa energia. Por sua vez, como estes não correspondiam a nada real, seu comportamento dificilmente seria linear, ou seja, a energia depositada não seria proporcional a energia incidente.

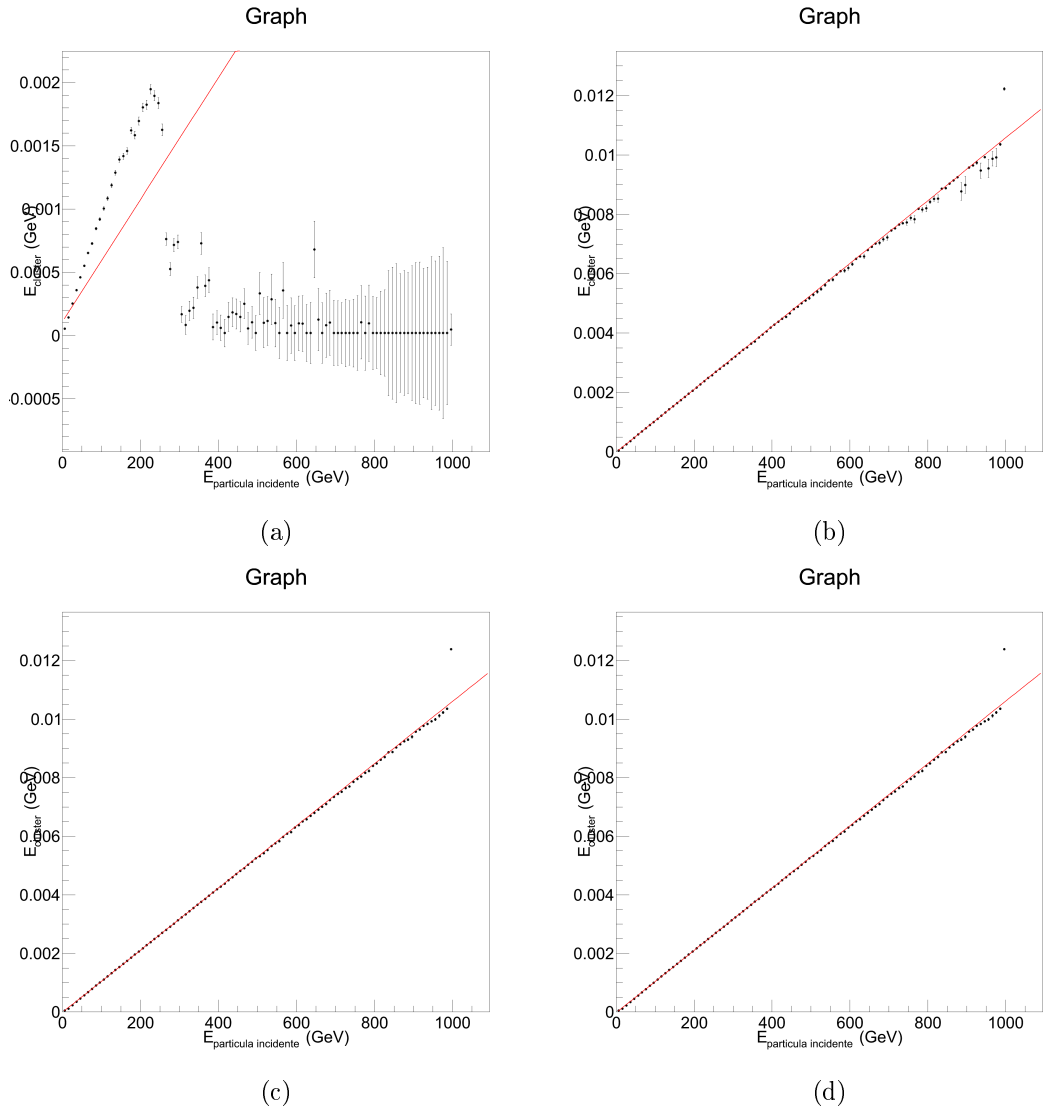


Figura 6.25: Curvas de calibração para os valores da tabela 6.2. (a) Corte 1. (b) Corte 2. (c) Corte 3. (d) Corte 4.

Na determinação de quais eventos utilizar para o cálculo, primeiro considerou-se como eventos válidos aqueles que tivessem pelo menos dois clusters encontrados nas camadas de alta granularidade (segmentos 1 e 3). Caso houvesse mais de dois, escolhia-se os mais energéticos, pois assumiu-se que estes representariam os fótons de decaimento, enquanto

os outros seriam artefatos do algoritmo. Após isso estes foram utilizados para o cálculo da massa invariante utilizando a equação 6.4.

$$m_\pi = \frac{E_{calib}^2}{2}(1-\alpha^2)(1-\cos\phi) \tag{6.4}$$

em que E_{calib} é a energia calibrada, ϕ é o ângulo entre as duas partículas filhas e

$$\alpha = \frac{E_{cluster1} - E_{cluster2}}{E_{cluster1} + E_{cluster2}} \tag{6.5}$$

α representa a assimetria de energia entre os dois clusters. E_{calib} é calculado tomando a energia depositada nos segmentos compostos por células de maior granularidade e relacionando com a energia da partícula incial através dos ajuste da figura 6.25. α é calculado utilizando a energia depositada nos segmentos de menor granularidade. ϕ é calculado utilizando o produto escalar entre os vetores definidos da origem (ponto de interação) até os pontos encontrados dos cluster no HGL.

Com isto se preencheu um histograma com as massas calculadas. Foram projetados todos os valores abaixo de 300 GeV (valores onde a massa calculada foi a mais próxima da real) e assim obteve-se uma gaussiana utilizada para criar uma faixa aceitável de massas invariantes. Estes resultados, para o corte 2, estão apresentados na figura 6.26. Neste gráfico é possível observar um reflexo da escolha dos raios de busca da seção 6.3.3, pois com o raio escolhido (0.3 cm), a faixa em que a massa invariante do píon é bem definida é a região de 300 GeV. Após a projeção no eixo y obteve-se os gráficos da coluna da direita na figura 6.26. Depois disso ajustou-se esta distribuiçã para uma gaussiana, obtendo assim sua média (M) e desvio padrão (SD). Com estes valores, definiu-se um intervalo de aceitância de massas invariantes ($[M-3SD, M+3SD]$).

Tendo estes intervalo recalculou-se as massas invariantes de todos os clusters utilizando novamente a fórmula 6.4. Foi observado se os resultados obtidos neste cálculo estavam dentro do intervalo de massas invariantes. Caso estivessem, era considerado que o algoritmo conseguiu identificar os fótons do decaimento de π^0. Utilizando este novo critério, obteve-se o gráfico para eficiência na identificação de π^0, utilizando a combinação de todos os segmentos, na figura 6.27.

Nestes gráficos pode-se observar uma grande melhora nas eficiências calculadas para a identificação de π^0. Deve-se fazer uma ressalva para o corte 1 (figura 6.27(a)). Apesar de ele aparentar um resultado bom, isto deve ser analisado com mais cuidado. Ao observarmos a figura 6.25(a) vemos que a resposta para este detector não é linear. Isto deve-se ao corte ter sido muito baixo, o que levou a inclusão de muitos clusters que não estão relacionados aos fótons que entraram no detector. Com isso o resultado obtido não pode ser tomado como correto apesar dele se comportar de maneira parecida com os outros cortes. Assim, este corte deve ser descartado. No caso do corte 4 (figura 6.27(d)), estes

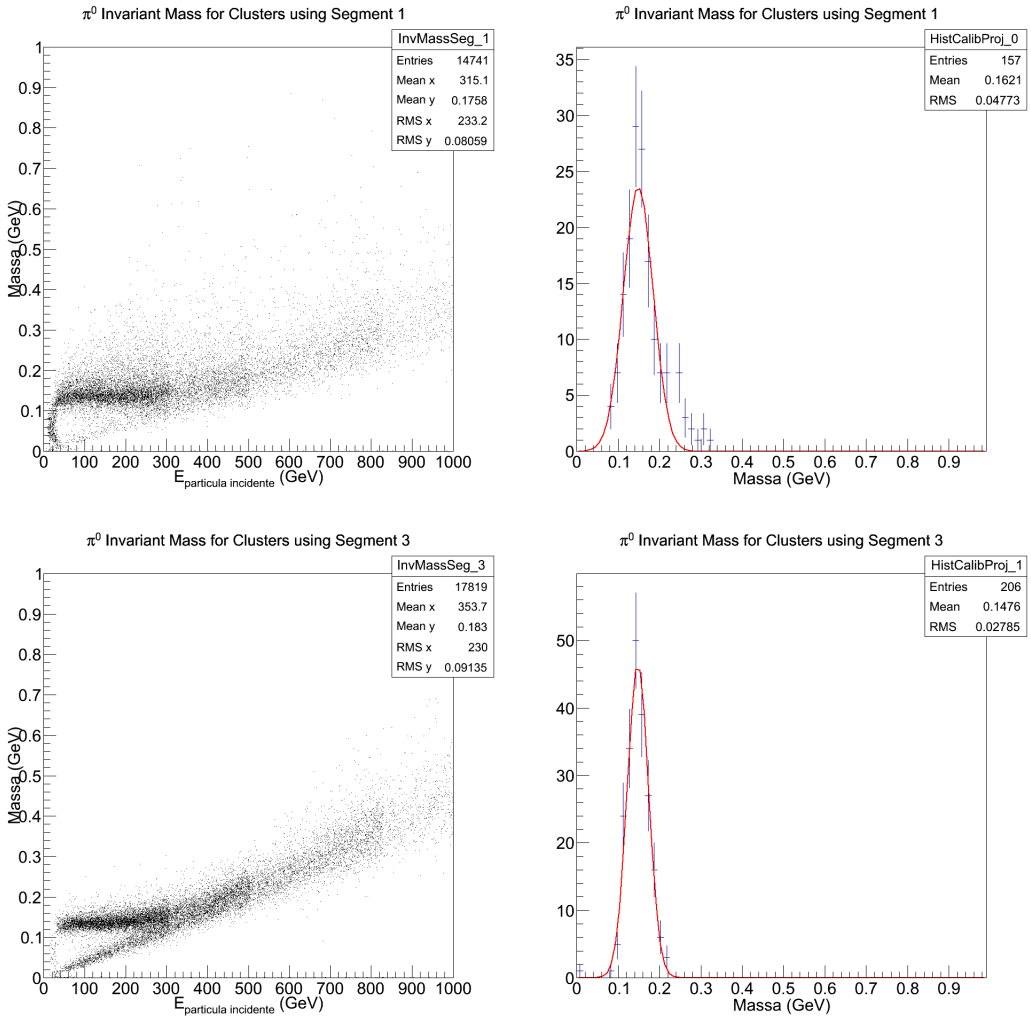
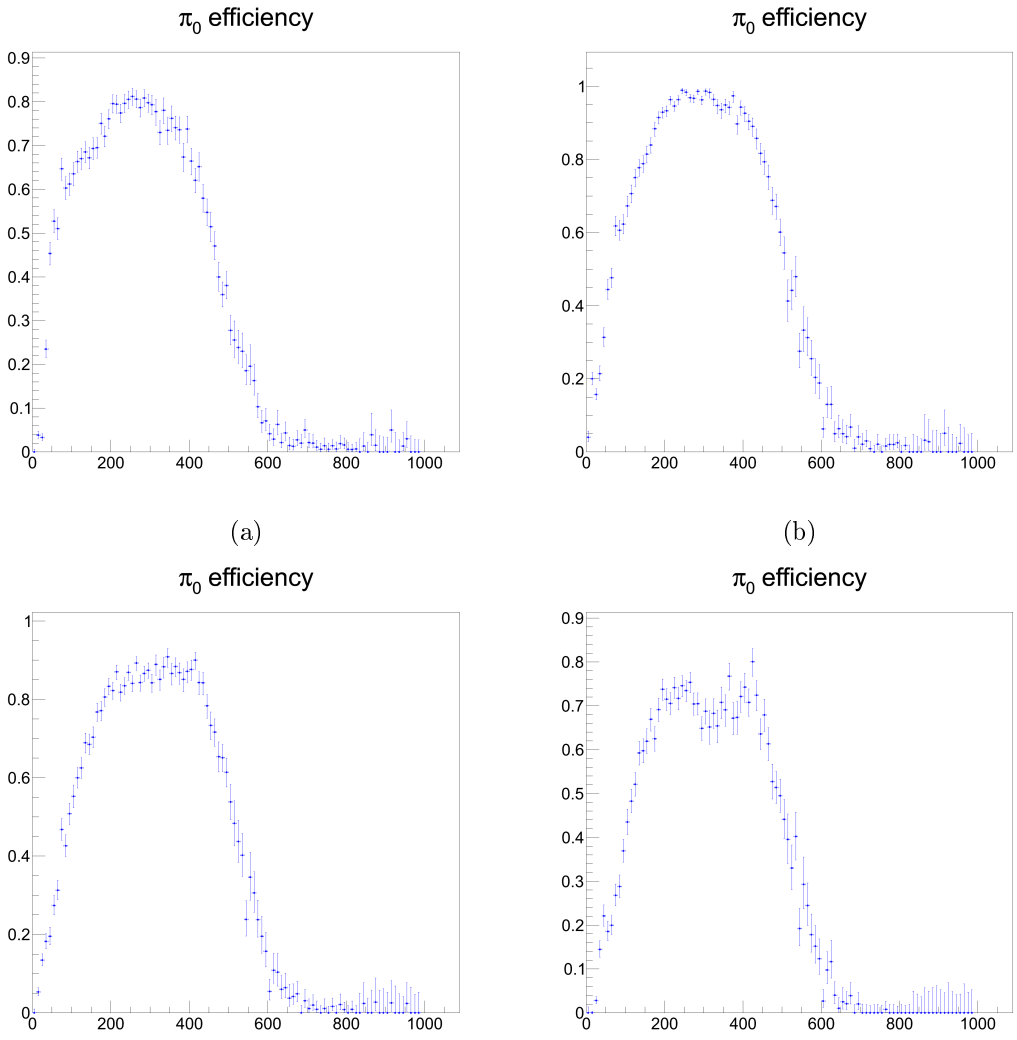


Figura 6.26: Histograma de massa invariante e seu ajuste para os segmentos 1 e 3.



(a) (b)

(c) (d)

Figura 6.27: Eficiência para identificação de π^0 pela energia da partícula incidente (GeV) recalculada, utilizando a combinação de todos os segmentos, para os quatros cortes da tabela 6.2. (a) Corte 1. (b) Corte 2. (c) Corte 3. (d) Corte 4.

foram muito severos, o que levou a perda de clusters reais (provenientes de fótons de π^0), diminuindo assim sua eficiência. Os melhores cortes foram os 2 (figura 6.27(b)) e 3 (figura 6.27(c)), pois estes não foram nem muito altos e nem muito baixos. Portanto, o corte ideal para este algoritmo estaria entre os valores dos cortes 2 e 3.

É possível, ainda, notar na figura 6.27 outro reflexo da limitação imposta ao raio de busca na seção 6.3.3. Relembrando, para o HGL, foi escolhido um raio de busca de 0.3 cm. Este valor correspondia a energias na faixa de 300 GeV a no máximo 500 GeV (figura 6.21). Assim neste gráfico é possível ver que a partir de 500 GeV a eficiência começar a diminuir devido ao tamanho imposto ao raio de busca, que torna a identificação de clusters com energia superior impossível. A eficiência na região entre 0 e 200 GeV é menor, pois os fótons de decaimento não atingem o detector devido aos seus grandes ângulos de separação na faixa de energia menor que 50 GeV. Acima desta faixa poderia ser atribuída essa baixa eficiência ao tamanho do raio de busca. Como ele é muito pequeno os clusters encontrados não contém toda energia depositada pelo chuveiro, levando a uma divisão deste em dois ou mais clusters, sendo que os falsos, geralmente, são criados na periferia do chuveiro. Isto leva, assim, a uma diminuição desta eficiência.

Capítulo 7

Conclusão

Neste trabalho foi realizado um estudo do comportamento de um algoritmo de clusterização para um calorímetro frontal. Foram realizados testes do algoritmo implementado por Taku Gunji para avaliar seu comportamento em geometrias diferentes. Após este período o algoritmo sofreu alterações para se adaptar às novas geometrias e para melhorar seu desempenho. Como o objetivo principal era a identificação de fótons de decaimento do π^0, este trabalho caminhou a este objetivo a partir da seção 6.3 e culmina com o cálculo da eficiência para a identificação de fótons do decaimento do π^0.

As limitações de ordem física (tamanho dos sensores) e de ordem lógica (ser do tipo *hard cluster*) refletem um desempenho baixo do algoritmo apresentado neste trabalho. O algoritmo consegue distinguir dois clusters com uma eficiência próxima de 100% numa faixa de 150 a 500 GeV para os valores de rapidez analisados ($y = 2.5, 3.0, 3.5, 4.0$), entretanto esperava-se uma eficiência próxima de 100% de 0 a 600 GeV. Esta limitação ocorreu devido à simplicidade do algoritmo e por ser do tipo *hard cluster*. Os algoritmos deste tipo tendem a ter uma fronteira com um tamanho fixo, não se adaptando a situações que exigiriam uma maior flexibilidade. Dada esta rigidez, no caso de chuveiros muito largos, parte da energia depositada acaba não sendo computada, levando a uma medida errada. No caso em que os chuveiros estão sobrepostos, a energia medida em cada cluster formado é subdimensionada. Como se está utilizando uma fronteira fixa, a energia depositada na parte sobreposta dos chuveiros não é computada adequadamente, pois nesta região a contribuição da energia provém de mais de um chuveiro. Isto acaba levando a erros na energia de cada fóton.

Uma solução para esta incapacidade de indentificação seria o uso de algoritmos baseados no método de Fuzzy, pois sua definição de sua fronteira não é rígida. Neste caso, os pontos que se encontram nesta fronteira podem pertencer a mais de um cluster. Isto é possível pois é dado um grau de "pertencimento" a cada cluster. Após o término da análise apresentada neste trabalho, um algoritmo baseado neste método esta sendo utilizado pelo colaboração do FoCal.

Apêndice A

Detectores de pixel

A noção de pixel (abreviação para *picture element*) foi introduzida em processamento de imagem para descrever a menor unidade discernível num elemento ou dispositivo. Um detector de pixel é um aparelho capaz de detectar uma imagem, e o tamanho do pixel é a granularidade da mesma. As câmaras digitais são exemplos de detectores de pixel. Neste caso, fótons de diferentes energias são integrados no sensores (pixel) durante o pequeno tempo de exposição e gerando uma distribuição de intensidade que é a imagem[33].

Um esquema de detector de pixel está apresentado na figura A.1. Nesta figura a partícula ionizante atravessa o sensor, gerando carga que, movendo-se na região de depleção sob ação de um campo elétrico, produz sinal. Estes são amplificados, e os pixels acionados são identificados e armazenados pela eletrônica. Esta figura é um exemplo de um tipo particular de detector de pixel: o detector de pixel híbrido[33].

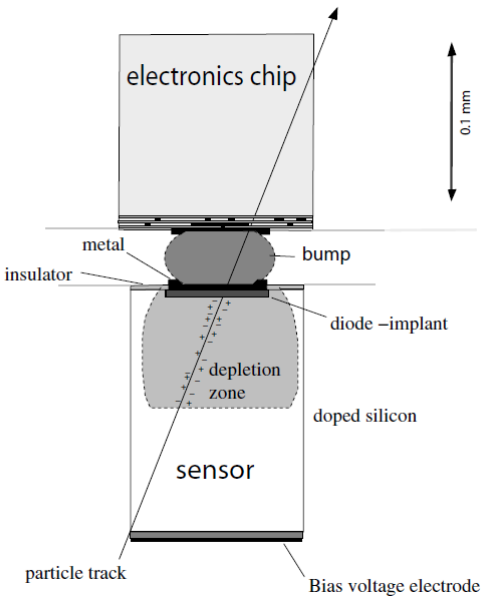


Figura A.1: Esquema de um sensor de pixel. Retirado de [33]

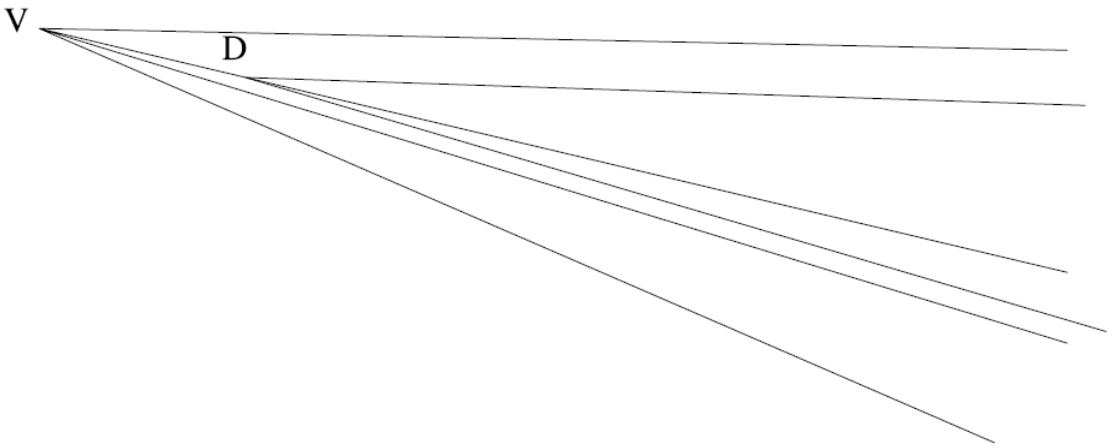


Figura A.2: Topologia de um decaimento de uma partícula com vida curta com outras partículas. O vértice da colisão (V) e o vértice de decaimento (D) estão indicados. Eles têm uma separação de alguns milímetros.

Aplicações destes detectores em física de partículas exigem que eles tenham uma boa resolução de tempo, trabalhem em altas taxas e tenham a habilidade de identificar padrões de pixels acionados (*hit patterns*). Em outras áreas, a ênfase é maior na sensibilidade e estabilidade[33].

Motivação para detectores de pixel em física de partículas

O desenvolvimento do detector de pixel em física de partículas tem sido motivado por dois requerimentos específicos, os quais se tornaram recentemente importantes e tem que ser simultaneamente alcançados:

(a) A possibilidade de estudar partículas com um tempo de vida curto;

(b) A capacidade de se ajustar com as altas taxas de interação e energia de modernos aceleradores de partículas.

Cientistas foram confrontados como seguinte problema: aceleradores de alta energia geram partículas elementares numa taxa de $10-100MHz$, com $10-100$ partículas sendo criadas a cada colisão. Algumas partículas raras vivem algo em torno de $1ps$ ($10^{-12}s$) e então decaem em algumas filhas. Um exemplo de topologia deste decaimento é mostrado na figura A.2. As trajetórias geradas neste decaimento devem ser medidas o mais próximo possível do ponto de interação[33].

A precisão não é o único parâmetro importante, pois muitas outras partículas podem passar próximas ao ponto de decaimento e isto pode interferir no seu local exato. Isto torna difícil ou impossível estudar o decaimento mesmo que um detector tenha a precisão

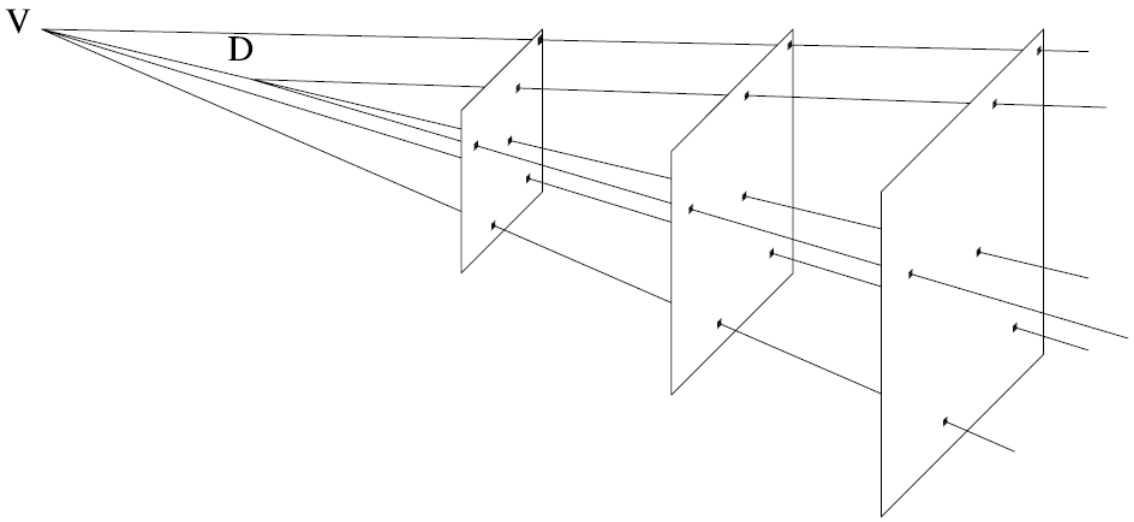


Figura A.3: Mesmo decaimento da figura A.2. As trajetórias são detectadas por três detectores de pixel. Os pixels acionados (o padrão visto pelo detector) estão marcados em preto.

necessária, a não ser que tenha elementos sensíveis suficientes. Os detectores de pixel têm uma resolução espacial e temporal suficiente, como também uma alta granularidade, resolvendo assim o problema anterior [33]. Esta afirmação pode ser ilustrada pelo seguinte exemplo.

Imagine um detector de pixel com quadrados de $0.1 \times 0.1 \quad mm^2$ cobertos por uma fina camada de silício. Assume-se que cada quadrado seja um sensor independente capaz de detectar a passagem da partícula. Se cada sensor é 100% eficiente e não há nenhum ruído, a figura A.2 aparece como um telescópio constituído de três detectores como mostrado na figura A.3.

Agora imagine que esses sensores sejam substituídos por outros com a mesma resolução temporal e espacial, mas que sejam ao longo de uma única coordenada. Isto reduziria drasticamente o número de canais e suas ligações seriam reduzidas drasticamente. Este padrão de sensores acionados é mostrado na figura A.4. Assume-se que a eficiência seja de 100% e que não há ruído.

Este é o primeiro tipo de detector utilizado em física de partículas que atacou os problemas gerados pelos requisitos (a) e (b). Neste caso a resolução desejada é alcançada utilizando uma grade de *microstrip*. Juntamente com as N *strips* acionadas pelas N trajetórias, deve-se levar em conta as $(N^2 - N)$ coincidências que ocorreram por engano[33]. Estas últimas ocorrem quando uma segunda partícula aciona duas strips e estas também cruzam com outras previamente acionadas. Com isso em vez de termos dois pontos (representando as acionadas pelas partículas), teremos 4 pontos (dois reais e dois fictícios). Esta coincidências podem afetar a reconstrução da trajetória.

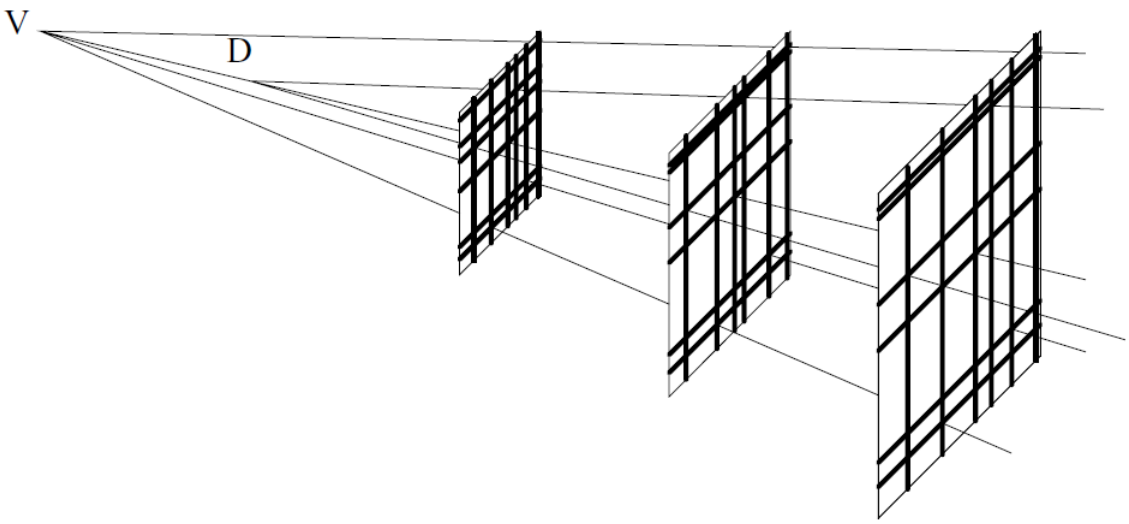


Figura A.4: Mesmo decaimento da figura A.2. As trajetórias foram medidas por três detectores de *strip* duplas. As *strips* acionadas (o padrão visto pelo detector) estão destacadas em preto.

Em resumo, os requerimentos (a) e (b) para física de partículas, são satisfeitos por um detector de alta granularidade capaz de registrar várias trajetórias com uma boa resolução temporal e espacial. Sua eletrônica deve ser capaz de selecionar os eventos de interesse, sendo que estes podem ser raros. Para fazer isso a eletrônica deve ser tal que possa temporariamente armazenar o padrão dos pixels acionados pertencentes a um dado evento. Estas variáveis são digitalizadas e serão usadas para selecionar (ou "*trigger*") os eventos de interesse. Cortes sucessivos são aplicados aos eventos até que o número destes seja reduzido e assim transferidos a um computador para análise[33].

Aplicações destes detectores para física de partículas requer que um padrão de *hits* (acionamento/imagem) seja um evento. A informação não é uniforme em todos os padrões possíveis, mas há alguns casos raros em que deve-se procurá-los através de algoritimos sofisticados (por exemplo clusterização) e eletrônica apropriada[33].

Detector de Pixel Híbrido

A fabricação de um sensor de pixel híbrido consiste em produzir o sensor e a unidade de processamento separadamente, e em alguns casos por métodos diferentes, e depois uni-los em um único sensor. A mudança no design do sensor (de tiras para pixels) tem várias consequências a nível de sistema e oferece várias aplicações [33]. Na figura A.5 é apresentada a visão explodida de um detector de pixel híbrido.

Fotos do detector e sua eletrônica usado como o primeiro detector de pixel híbrido aplicado nos experimentos de física de altas energias está na figura A.6.

O detector de pixel hibrido é um detector ideal para trabalhar em ambientes hostis

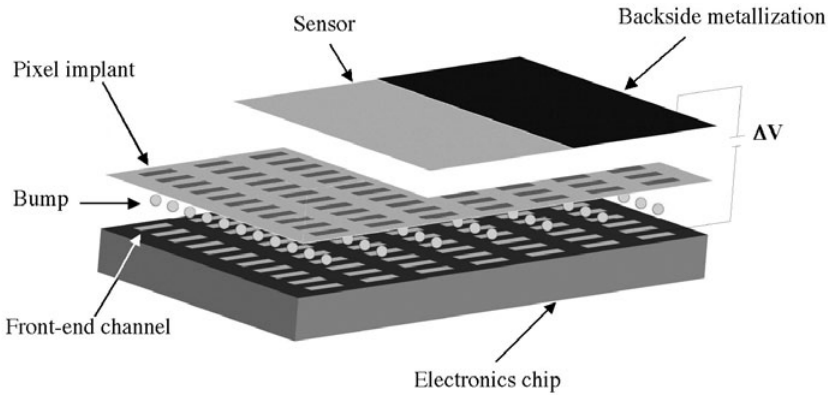


Figura A.5: Esquema da visão explodida de um detector de pixel híbrido[33].

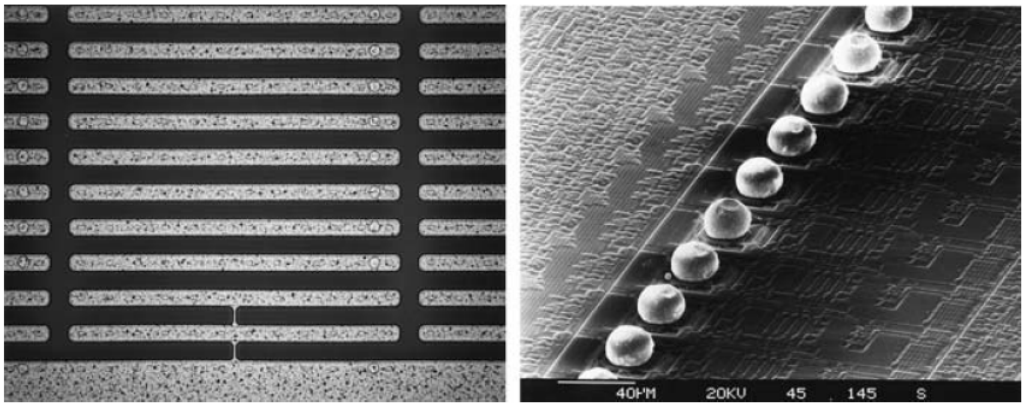


Figura A.6: Foto do primeiro detector de pixel híbrido usado em experimentos de física de altas energias. As estruturas interessantes de ambos detector e eletrônica não são visíveis depois da montagem. Na figura da esquerda está apresentado um *close* do detector. Na figura da direita foi dado um *zoom* dos *bumps* de ligação e da eletrônica. [34]

próximo a região de interação de um acelerador de partículas porque:

- é resistente a radiação (suporta um fluxo intenso de partículas);
- é capaz de fazer medidas tri-dimensionais com uma boa resolução temporal;
- Fornece resolução espacial para estudar partículas de vida curta;
- Pode extrair raros padrões que são de interesses dos físicos (memoriza o padrão dos pixels acionados e seleciona aqueles de interesse).

Esta tecnologia tem sido a escolha para a próxima geração de detectores em física de altas energias no LHC e em experimentos de alvo fixo. Esta escolha foi tomada pelo fato de que esta tecnologia é a única atualmente madura para se construir detectores com áreas maiores que alguns cm^2. As principais vantagens desta tecnologia para detectores são as seguintes[33]:

- Este sistema permite o teste em vários passos intermediários, oferecendo assim um alto rendimento na produção de módulos com áreas de dezenas de cm^2.
- O chip, sensor, e a tecnologia de interconexão são processos industriais utilizados há vários anos. Eles estão disponíveis em uma variedade de industrias.
- Devido ao fato do sensor e o chip serem itens separados, outros materiais além do silício podem ser usados como substrato para o sensor.

As desvantagens do detector híbrido se tornam evidente quando é necessário uma alta resolução num ambiente de alta multiplicidade e altas velocidades. As limitações tecnológicas são principalmente relacionadas à união entre o sensor e chip, a densidade de energia associada a limitação dos circuitos eletrônicos para amplificação e lógica confinados na mesma área do eletrodo detector[33].

Detectores de Pixel Monolíticos (*Monolithic Pixel Detectors*)

Como o silício é o material mais comumente usado para detectores de pixel, vários grupos têm estudado a possibilidade de se construir o sensor e a eletrônica num mesmo processo. Isto evita a grande densidade de conexões e outras manipulações relacionadas. Isto abre a possibilidade para um detector mais robusto e barato, mas menos versátil[33].

Em alguns casos o tipo de sensor é que determina qual será o desenvolvimento do detector. Circuitos elétricos simples (primeiro estágio de amplificação de cada pixel e alguns esquemas de endereços) são integrados em silício de alta resistividade. A geração do sinal é ótima devido a grande região de depleção que pode ser criada e a altos campo elétricos que fornecem uma rápida e eficiente coleta das cargas. O tratamento do sinal no

chip é mínimo porque todas as tecnologias têm sido desenvolvidas para silício com baixa resistividade e um considerável esforço tem sido gasto em criar circuitos simples com um rendimento aceitável[33].

Em outros casos a eletrônica é a que determina o desenvolvimento do detector. O detector é colocado numa fina camada de silício dopado do tipo p de baixa resistividade, o que é ótimo para o desenvolvimento de circuitos complexos mas não permite uma zona depleção grande, o que não proporciona uma rápida coleta de cargas.

Nos anos iniciais dos detectores semicondutores a integração monolítica utilizada em larga escala era vista como o "Santo Graal". Claramente, é um conceito atraente ter um detector de $6 \times 6cm^2$ que combina um detector de *strip* e 1200 canais de leitura com somente a alimentação e a leitura dos dados como conexão externa. O problema foi percebido quando notou-se a incompatibilidade entre processo de fabricação do CI (circuito integrado) e do detector. O desenvolvimento de um CI compatível com o processo do detector permitiria uma integração monolítica de alta qualidade e sensor de silício completamente tomado pela região de depleção sem perda na sua performance. Porém uma simples estimativa do rendimento mostra que isto não é prático, pois como são dispositivos complexos sua eficiência não é de 100%[35].

Alguns design de sensores de pixels copiam completamente o sistema de leitura usado para CCDs. Este sistema é uma boa combinação para a fotografia digital, em que cada pixel carrega as informações necessárias. O circuito eletrônico para este sistema é bastante simples e a leitura pode ser lenta. Leituras lentas para a detecção de partículas carregadas também permite circuitos simples e baixo ruído elétrico. Contudo, em ambientes com dados esparsos com altas taxas de eventos (por exemplo em colisores de hádrons de alta luminosidade) requer uma rápida resposta, pois precisa-se determinar quais pixels foram acionados e armazenar esta informação temporariamente, levando a aumento da complexidade do circuito[35].

Apêndice B

Rapidez

Em colisões de íons pesados relativísticos e em muitos outros processos de altas energias é conveniente usar variáveis cinemáticas que tem propriedades simples quando se muda o sistema de referência. A rapidez y é umas destas variáveis, juntamente com a pseudorapidez η e as variáveis de cone de luz. Todas elas são variáveis cinemáticas que têm propriedades simples durante uma transformada de Lorentz. Neste apêndice trataremos somente da rapidez.

Antes, iremos definir algumas notações e convenções. Iremos usar as coordenadas naturais $c = \hbar = 1$. As coordenadas são definidas por um vetor contravariante com componentes x^μ[36]:

$$x^\mu = (x^0, x^1, x^2, x^3) = (t, \mathbf{x}) = (t, x, y, z). \tag{B.1}$$

O vetor momento também é definido como um vetor contravariante com componentes p^μ:

$$p^\mu = (p^0, p^1, p^2, p^3) = (E, \mathbf{p}) = (E, \mathbf{p_T}, p_z) = (E, p_x, p_y, p_z). \tag{B.2}$$

A rapidez de uma partícula é definida em termos de suas componentes energia e momento p_0 e p_z por

$$y = \frac{1}{2} \ln \left(\frac{p_0 + p_z}{p_0 - p_z} \right). \tag{B.3}$$

Esta é uma quantidade adimensional. No limite não relativístico a rapidez de uma partícula viajando na direção longitudinal é igual a velocidade da partícula em unidades da velocidade da luz. A rapidez depende do sistema de referência, mas se transforma de uma maneira simples. A rapidez em um sistema é relacionada a de outro através de uma constante aditiva[36].

Da definição B.3, temos que

$$e^y = \sqrt{\frac{p_0 + p_z}{p_0 - p_z}} \tag{B.4}$$

e

$$e^{-y} = \sqrt{\frac{p_0 - p_z}{p_0 + p_z}} \tag{B.5}$$

Somando as equações B.4 e B.5 temos a relação entre a energia p_0 e a rapidez y:

$$p_0 = m_T cosh y, \tag{B.6}$$

em que m_T é a massa transversal da partícula:

$$m_T^2 = m^2 + \mathbf{p_T}^2 \tag{B.7}$$

subtraindo as equações B.4 e B.5 temos a relação entre o momento longitudinal p_z e a rapidez y:

$$p_z = m_T senh y, \tag{B.8}$$

As equações B.6 e B.8 são úteis para relacionar as componentes do momento com a variável rapidez[36].

Nós observamos que sobre uma transformação de Lorentz do sistema do laboratório F para um novo sistema F' se movendo com velocidade β na direção z, a rapidez y' da partícula no sistema de referência F' é relacionada a rapidez y do sistema antigo F por

$$y' = y - y_\beta \tag{B.9}$$

em que y_β é a rapidez que a partícula teria no sistema F, se ela estivesse com a velocidade β do sistema em movimento F' e é dada por

$$y_\beta = \frac{1}{2} \ln \left(\frac{1+\beta}{1-\beta} \right). \tag{B.10}$$

Esta quantidade pode ser chamada de "rapidez do sistema de referência em movimento". Assim a rapidez de uma partícula no sistema em movimento é igual a rapidez do sistema em repouso menos a rapidez do sistema em movimento, similar a subtração de velocidades do sistema em movimento do caso não relativístico. É frequentemente útil tratar a rapidez como uma medida relativística da "velocidade" da partícula. Esta propriedade simples sob uma transformação de Lorentz a faz uma escolha boa para tratar a dinâmica de partículas relativísticas. Outra informação importante sobre esta variável é que dada uma energia incidente, a rapidez da partícula incidente e do alvo podem ser determinadas facilmente e que quanto maior a energia, maior é a separação entre a rapidez da partícula incidente e a rapidez da partícula alvo[36].

Referências Bibliográficas

[1] ATLAS Collaboration, Physics Letters B **688**, 21 (2010).

[2] CMS Collaboration, Journal of High Energy Physics **2010**, 1 (2010), 10.1007/JHEP02(2010)041.

[3] LHCb Collaboration, Physics Letters B **693**, 69 (2010).

[4] ALICE Collaboration, Eur. Phys. J. C **65**, 111 (2010).

[5] L. Evans, Eur. Phys. J. C **34**, 57 (2004).

[6] ALICE Collaboration et al., Journal of Physics G: Nuclear and Particle Physics **30**, 1517 (2004).

[7] ALICE Collaboration et al., Journal of Physics G: Nuclear and Particle Physics **32**, 1295 (2006).

[8] Ahmad et al., Alice technical proposal, Technical report, CERN, 1995.

[9] C. Grupen and B. A. Shwartz, *Particle Detectors*, Cambridge University Press, 2008.

[10] R. Wigmans, *Calorimetry*, International Series of Monography on Physics, Oxford University Press, 2008.

[11] D. Green, *The Physics of Particle Detectors*, volume 12 of *Cambridge monographs on particle physics, nuclear physics and cosmology*, Cambridge university press, 2000.

[12] *Techniques for Nuclear and Particle Physics Experiments - A How to Approach*, chapter 5. General Characteristics of Detectors, pages 111–112, Springer-Verlag.

[13] https://twiki.cern.ch/twiki/bin/view/ALICE/FoCAL .

[14] U. A. W. Paloma Quiroga-Arias, Jose Guilherme Milhano, arXiv:1002.2537v1 [hep-ph] .

[15] Y. Hori, Simulation study for forward tracking calorimeter in lhc-alice experiment, Master's thesis, Department of Physics, Graduate School of Science, University of Tokyo.

[16] E. A. Kuraev, L. N. Lipatov, and V. S. Fadin, Sov. Phys. JETP **45**, 199 (1977).

[17] I. I. Balitsky and L. N. Lipatov, Sov. J. Nucl. Phys. **28**, 822 (1978).

[18] Y. L. Dokshitser, Sov. Phys. JETP (1977).

[19] G. Altarelli and G. Parisi, Nuclear Physics B **126**, 298 (1977).

[20] D. d'Enterria and the CMS Collaboration, Journal of Physics G: Nuclear and Particle Physics **34**, S709 (2007).

[21] T. K. Nayak, arXiv:1009.2220v1 [nucl-ex] .

[22] L. McLerran, arXiv:0812.4989 [hep-ph] .

[23] http://geant4.web.cern.ch/geant4/support/userdocuments.shtml .

[24] http://aliceinfo.cern.ch/Offline/AliRoot/Manual.html .

[25] V. K. Pang-Ning Tan, Michael Steinbach, *Introduction to Data Mining*, Addison-Wesley, 2005.

[26] http://home.dei.polimi.it/matteucc/Clustering/ tutorial_html/index.html .

[27] A. K. Jain, Pattern Recogn. Lett. **31**, 651 (2010).

[28] N. J. Nilson, *Introduction to Machine Learning: An early draft of a proposed textbook*, Department of Computer Science, Standfor University, Standford CA 94305, 1998.

[29] J. F. Trevor Hastie, Robert Tibshirani, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer.

[30] B. F. János Abonyi, *Cluster Analysis for Data Mining and System Identification*, Die Deutsche Bibliothek.

[31] R. Jarvis and E. Patrick, IEEE Transactions on Computers , 1025 (1973).

[32] E. Parzen, The Annals of Mathematical Statistics **38**, 1065 (1962).

[33] N. W. et. al., *Pixel Detectors from Fundamentals to Application*, Particle Acceleration and Detection, Springer.

[34] G. Lutz, *Semiconductor Radiation Detector: Device Physics (Accelerator Physics)*, Springer Verlag, 1999.

[35] H. Spieler, *Semiconductor Detector System*, Number 12 in Series on Semiconductor Science and Technology, Oxford University Press, 2005.

[36] C.-Y. Wong, *Introduction to high-energy heavy-ion collisions*, chapter 2. Kinematics Variables, World Scientific Publishing Co. Pte. Ltd.