Nuclear Instruments and Methods in Physics Research A 425 (1999) 536–548

NUCLEAR INSTRUMENTS & METHODS IN PHYSICS RESEARCH Section A

Optimisation of clustering algorithms for the reconstruction of events started by a 1 GeV photon beam in a segmented BGO calorimeter

A. Zucchiatti[a,*], D. Moricciani[b], A.M. Massone[c], F. Masulli[c,d], M. Capogni[e], M. Castoldi[a], A. D'Angelo[b], F. Ghio[f], B. Girolami[f], P. Levi Sandri[g], M. Sanzone[a]

[a] *Istituto Nazionale di Fisica Nucleare Sezione di Genova, Via Dodecaneso 33, 16146 Genova, Italy*
[b] *Istituto Nazionale di Fisica Nucleare, Sezione di Roma II, Roma, Italy*
[c] *Istituto Nazionale per la Fisica della Materia, Genova, Italy*
[d] *DISI – Dipartimento di Informatica e Scienze dell'Informazione, Università di Genova, Genova, Italy*
[e] *Institut de Phisique Nucleaire, Orsay Cedex, France*
[f] *Istituto Superiore di Sanità and INFN sezione Roma III, Roma, Italy*
[g] *Istituto Nazionale di Fisica Nucleare, Laboratori Nazionali di Frascati, Frascati, Italy*

Received 24 September 1998

Abstract

Three different clustering algorithms have been implemented to reconstruct the response of a segmented BGO calorimeter to electromagnetic showers and hadrons. The ability of each algorithm to identify the number of interacting particles and to attribute to each particle the appropriate energy, has been assessed by comparison to various reactions simulated with the GEANT code. General considerations on the calorimeter response are made. A few significant reaction channels are discussed in detail as regards cluster identification and background reduction with each clustering technique. © 1999 Elsevier Science B.V. All rights reserved.

PACS: 29.40.V; 07.05T; 07.05.M; 07.85

Keywords: Calorimeters; BGO; Clusters; Simulation

1. Introduction

Segmented 4π calorimeters are arrays of a few to several thousand cells (detectors) that can provide, in great detail, a picture of nuclear reaction channels where several charged and neutral particles are involved in the final state. Photons and electrons above a few MeV develop, in interacting with the calorimeter medium, an electromagnetic shower that can exceed the physical boundary of a single cell and therefore produce a signal in a few of them grouped around the particle trajectory. Hadrons

[*] Corresponding author. Tel.: + 39 010 353 6384; fax: + 39 010 313 358; e-mail: zucc@ge.infn.it.

PII: S0168-9002(98)01402-8

can as well hit more than a single cell even if they produce a shower only at very high energy. The multiplicity associated to each particle, i.e. the number of cells that give a signal, depends on the particle type, energy, granularity (segmentation) of the calorimeter, shape of the cells and intrinsic properties of the material (Molière radius). One of the essential steps in the extraction of physical information from a segmented calorimeter is the identification and association of groups of cells related to a single interacting particle. This can be done by clustering algorithms of various nature. Only after clustering has been performed one can determine the number of particles detected by the calorimeter, their energy and, by centre of mass methods, their flight direction, thus describing the reaction final state. One could also associate the multiplicity and topology of the cluster to the particle type but in this work we have dealt only with cluster recognition. In 4π calorimeters, a few possible sources of data misinterpretation must be considered. Some secondary photon from a developing shower could be emitted at large angles from the direction of the primary particle, could cross the central area of the detectors assembly and re-enter the calorimeter giving rise to a secondary satellite cluster, in time with the experiment trigger. Some secondary neutral particles could be produced in a cell and not interact in the neighbour cells but further away, without crossing the central area of the detector assembly, thus breaking the contiguity of energy release and creating an energy spill. In several cases, strictly related to the nature of the reaction channel observed and to the granularity of the calorimeter in use, two or more individual particles could begin their interaction in near cells and the showers they individually develop could merge to some extent. The ability of a clustering algorithm to disentangle critical cases and produce the correct number of primary particles emitted in the reaction is to be evaluated in all those cases were the detection efficiency is crucial to the extraction of physical information.

2. The simulation of events

It is evident that only events generated by a simulation code, bear all the information necessary to assess the efficiency of cluster reconstruction allowed by the different methods, since in real events the number of detected particles is a priori unknown. We have produced simulated data with the GEANT code [1] associated to the FLUKA [2] hadron interface. The GRAAL BGO array [3], which we have considered in this study together with its ancillary detectors, includes 480 cells (24 cm long crystals), covering the entire ϕ angle with 32 detectors having a $\Delta\phi$ angle of 11.25° and covering the θ angle from 25° to 155° with 15 detectors having a $\Delta\theta$ angle from 6.25° to 10.5°. We have not simulated the response of our calorimeter to a variable number of electromagnetic probes or hadrons randomly dispersed over the detector volume. Instead we have preferred to consider the behaviour of the detectors in real conditions, thus comparing the performance of different clustering methods on events that reproduce the multiplicity, the segmentation and the superposition of clusters, as they are determined by the kinematics of the observed reactions and will be found experimentally. Nuclear reactions, initiated on the proton by a high energy ($300\ \mathrm{MeV} < E\gamma < 1100\ \mathrm{MeV}$) γ-ray, extracted according to a Compton back-scattering distribution [4], have been produced by an event generator [5], either selectively, or according to their cross sections. To control the response of the calorimeter to each particle we have exploited one of the features of the GEANT package, that allows to rearrange the particle stack, by choosing and saving a non-standard value of the parameter UPWGHT. For an event let us say that k primary particles have been generated in the target, depending on the reaction channel chosen by the event generator. Decay products from short-lived primary particles, practically generated at the reaction point, have been considered amongst the k primary particles. In standard procedure all particles are automatically assigned a UPWGHT value of 1 and are pooled from the stack in a FIFO way. In our case each primary particle has been tracked through the different media (target cell, wire chambers, plastic scintillators, etc.) up to the step where it entered the electromagnetic calorimeter. At this stage it has been marked assigning the value $\mathrm{UPWGHT} = -k*100$, stored in the particles stack with its momentum, energy and entrance point co-ordinates and stopped. Obviously this procedure ends when all k primary particles have entered the electromagnetic calorimeter or escaped from the detector assembly. From there onwards the primary particles are pooled from the stack in descending UPWGHT order: -100, $-200,\dots,$ $-k*100$. Secondary particles, generated in the BGO are tracked as well. If their trajectory directs them out of the BGO, towards the target area and then back into the BGO, they are tagged by the value UPWGHT $= -101$, -102, etc or -201, -202 etc, according to the index of the primary particle that generated them. Since the UPWGHT value marks uniquely each particle, it is possible to analyse the entire event map (the location of all active cells) in terms of each particle contribution (the location of cells made active by that particle) and to know, in subsequent analysis, to which primary particle each secondary belongs.

3. Results of the simulation

We have examined the calorimeter response in the most general case; i.e. when all possible reaction channels are considered, weighted by their cross section. We have also examined three selected reaction channels of particular interest:

$$\gamma + \mathrm{p} \rightarrow \eta\mathrm{p} \rightarrow \eta\pi^0\pi^0\pi^0\mathrm{p}, \quad \gamma + \mathrm{p} \rightarrow \eta\mathrm{p} \rightarrow \pi^0\gamma\gamma\mathrm{p}$$

and

$$\gamma + \mathrm{p} \rightarrow \pi^0\mathrm{p} \rightarrow \gamma\gamma\mathrm{p}.$$

We have looked at:

1. the generation and rejection of primary and secondary particles,
2. the energy spills: i.e. single particle maps whose crystals are not all adjacent,
3. the merging clusters: i.e. maps belonging to different particles that have some adjacent elements and could therefore be seen as a unique compact cluster,
4. the distorted clusters: i.e. compact maps generated by a single particle but characterised by more relative maxima.

All quantities have been studied as a function of the E_{th} threshold, set uniquely for all 480 crystals, and as a function of the selected reaction channel. The E_{th} value has been changed from 1.8 MeV, corresponding to the hardware noise level, up to 15 MeV.

The efficiency of the calorimeter in detecting primary and secondary particles is evaluated from the ratio between the number of particles giving a signal above E_{th} in at least a crystal and the number of particles entering the calorimeter active volume. As shown in Fig. 1a the fraction of detected primary particles remains above 89% when all reaction channels are considered and is higher than 95% for

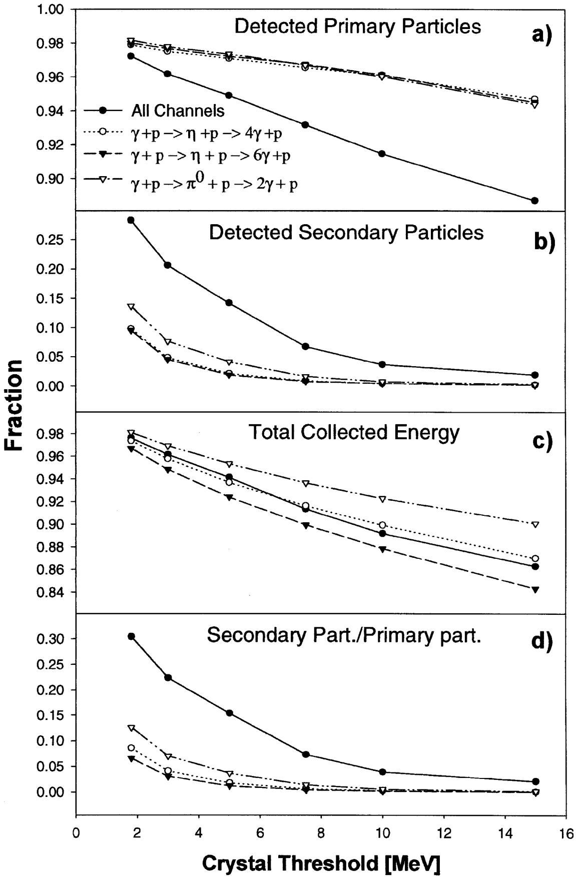

Fig. 1. The fraction of primary particles (part a), of secondary particles (part b), and of total energy (c) detected by the calorimeter as a function of the energy threshold set for the 480 BGO crystals. In part d is given the average number of secondary particles generated by each primary particle. All quantities are evaluated for different simulated reaction conditions.

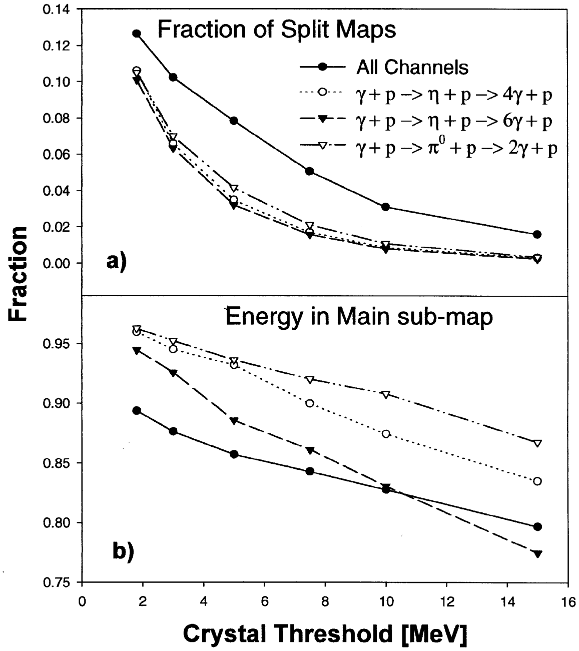

Fig. 2. The fraction of single particle maps that are split in more sub-maps (part a) and the fraction of energy deposited by the particle in the main sub-map (part b) as a function of the energy threshold set for the 480 BGO crystals. All quantities are evaluated for different simulated reaction conditions.

the three selected photoproduction channels. The secondary particles are much more sensitive to the threshold as seen in Fig. 1b. About 28% of the generated secondary particles are detected at $E_{\mathrm{th}} = 1.8$ MeV, reduce to about 7% at $E_{\mathrm{th}} = 7.5$ MeV and then drop slowly. For the three selected photoproduction channels the collection efficiency for secondary particles does never exceed 15% and is negligible above $E_{\mathrm{th}} = 7.5$ MeV. More similar in the four cases is the fraction of total energy collected, which goes from 98% at $E_{\mathrm{th}} = 1.8$ MeV to about 85% at $E_{\mathrm{th}} = 15$ MeV (Fig. 1c). On average a secondary particle is detected in association with a primary particle (Fig. 1d) in between 30% and 3% of the cases when all the reaction channels are considered. For the three selected photoproduction channels a secondary particle is detected in at most 10% of the cases but only in a 1–2% of the cases above 7.5 MeV.

There is a significant percentage of particles, both primary and secondary, that produce split maps on the calorimeter: i.e. maps where the crystals are not all adjacent. The fraction of split maps goes smoothly, as seen in Fig. 2a, from 13% at $E_{\mathrm{th}} = 1.8$ MeV down to 2% at $E_{\mathrm{th}} = 15$ MeV when all the reaction channels are considered. On average the number of sub-maps is in this case slightly above 2, with the major sub-map carrying between 80% and 90% of the energy associated to the particle, as seen in Fig. 2b. When the selected photoproduction channels are considered, the fraction of split maps (Fig. 2a) is still around 10% at low E_{th} but is only a few percent above $E_{\mathrm{th}} = 7.5$ MeV. The collected energy (Fig. 2b) is higher than in the more general case.

In Fig. 3 we have reported the situation concerning overlapping maps: i.e. maps belonging to different particles, that have a cell in common or have at least two adjacent cells. The fraction of overlapping maps is very sensitive to the selected reaction channel, as shown in Fig. 3a, being related to the number of primary particles in the final state. When all channels are considered, the dominant reactions are single pion photoproduction. In this case at most 7% of the particles produce maps merged with another one at $E_{\mathrm{th}} = 1.8$ MeV and not more than 0.9% at $E_{\mathrm{th}} = 15$ MeV. When the number of particles in the final state increases (Fig. 3a), the overlap probability does as well and the fraction of merged maps goes up to about 25% and remains above 4–6% at all thresholds. We have also computed the fraction of particles that generate maps whose maxima are in adjacent crystals: a situation where the separation appears impossible with the adopted clustering techniques. As seen in Fig. 3b this effect is not affected by the threshold since we know that more than 50% of the energy released by photons [6], electrons and hadrons is confined in a single crystal, increasing to 80% when the incident photon travels along the crystal axis. In the most general case this is the dominant effect above $E_{\mathrm{th}} = 7.5$ MeV but it affects only 1% of the cases. When several particles are present in the reaction final state, as in the three selected photoproduction channels, the percentage of cases that cannot be solved increases to 4–9%.

Finally we have computed, as shown in Fig. 4, the fraction of maps that are distorted; i.e. those maps that belong to a single particle, include only

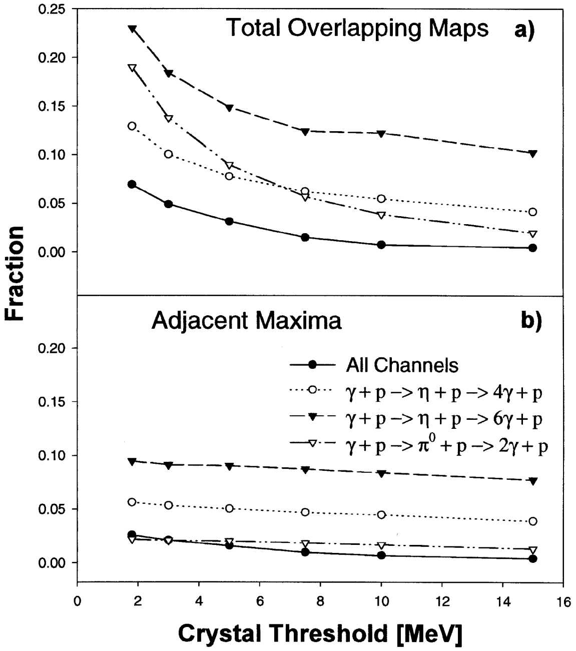

Fig. 3. Part a: the fraction of single-particle maps that overlap either by having coincident or adjacent maxima or by having coincident or adjacent crystals. Part b: the fraction of single-particle maps that overlap by having coincident or adjacent maxima. All quantities are evaluated as a function of the energy threshold set for the 480 BGO crystals and for different simulated reaction conditions.

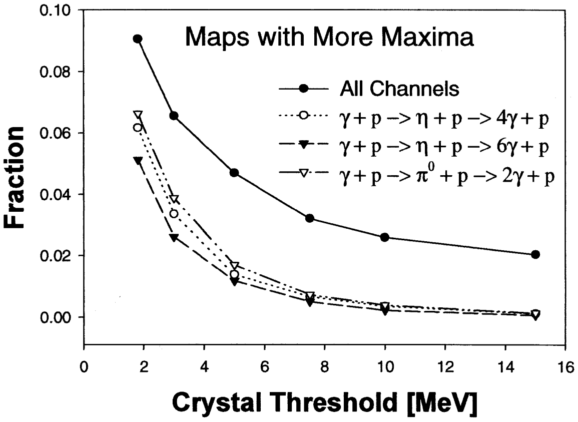

Fig. 4. The fraction of single-particle that produce distorted results showing more relative maxima in the map (if contiguous) or in the main sub-map (if split), as a function of the energy threshold set for the 480 BGO crystals and for different simulated reaction conditions.

adjacent crystals or, are split in more sub-maps, but have more than one relative maximum in the whole map (if contiguous) or in the major sub-map (if split). This effect is significant,in the most general case, only at low E_{th}, although in never exceeds 9%. For the selected photoproduction channels the percentage goes quickly below 1–2%.

All plots show a fast decrease up to $E_{th} = 5$ MeV and a slow trend above $E_{th} = 7.5$ MeV, with the exception of the fraction of particles producing maps with adjacent maxima. It appears that $E_{th} = 7.5$ MeV is a good compromise between the need of reducing the effect of secondary particles and the need of extracting with precision both the number of particles in the reaction final state and their energy.

4. Description of clustering algorithms

Our array can be viewed as a 15×32 BGO matrix in the θ–ϕ plane. We have developed and used three different algorithms amongst the many that have been proposed for similar instrumentation: a contiguity algorithm, a cellular automata method [7] and a deterministic annealing procedure [8]. In general terms any clustering procedure consists in taking elements from a set $X = \{\boldsymbol{x}_i, i = l, m\}$ defined by a vector $\boldsymbol{x}_i$ in a p-dimensional space and in associating each of them to a class taken from a set of classes $\boldsymbol{Y} = \{\boldsymbol{y}_j, j = l, n\}$ identified in the p-dimensional space by the coordinates $\boldsymbol{y}_j$ of their centroid. In the case of an electromagnetic calorimeter we operate in a three-dimensional space since we have for each cell $X_i = \{\theta_i, \phi_i, E_{cry}(i)\}$. However, there is no correlation between energy and position at the clustering stage, before the particle and the reaction are uniquely identified. Only the size of the shower $(\Delta\theta_i, \Delta\phi_i)$ might be related to its energy. Clustering will then consist in recognising groups of cells that are simply spatially associated. Nevertheless, we will use energy information to take decisions in ambiguous situations.

4.1. Contiguity method

This algorithm starts from a list of all crystals that have given a signal (active crystals), ordered first in increasing θ-index (from 1 to 15) and, for equal θ, in increasing ϕ-index (from 1 to 32) regardless of the energy deposited in each crystal. Beginning from the lowest (θ, ϕ) indexes, the algorithm checks if any of the adjacent crystals (5 for θ-index $= 1$ or 15, 8 elsewhere) is in the list of active crystals. If yes the neighbours are associated to cluster number 1 and the scan proceeds with further neighbours until a pattern of contiguous crystals can be followed. Then the first crystal in the list, that is not attributed to the cluster 1, is taken as the seed of cluster number 2 and the procedure of associating neighbours and further neighbours cells is repeated until contiguity can be maintained. When all active crystals are associated to a cluster the procedure stops. This is the simplest and fastest of all clustering methods but cannot avoid energy spills, satellite clusters and merging showers, unless the energy threshold E_{th} is increased to a level that is above the energy deposited in crystals belonging to a "spill", a "satellite" or above the valley region between two or more merging showers. This will lead to elimination of secondary effects and also to an evident generalised reduction in the energy attributed to each primary interacting particle.

4.2. Cellular automata

This algorithm is based on transition rules equivalent to those of Breton et al. [7]. To each active crystal i ($1 \leq i \leq m$) it is associated a structure of this kind:

1. *crystal.ite* is the θ-index of the crystal (from 1 to 15).
2. *crystal.ifi* is the ϕ-index of the crystal (from 1 to 32).
3. *crystal.ecry* is the energy $E_{\text{cry}}(i) > E_{\text{th}}$ deposited in the crystal.
4. *crystal.clclu* is the progressive number j ($1 \leq j \leq n$) of the cluster to which the crystal gets associated.
5. *crystal.fever* is the energy $E_{\text{cl}}(j)$ accumulated in the whole cluster j to which the crystal belongs.
6. *crystal.cnt* is an index normally equal to 0 and set to 1 when the cluster j, to which crystal i belongs, is contiguous to some other cluster.

The m structures are ordered, to begin, by descending E_{cry} values. This means that the first crystal in the list is the absolute maximum as regards energy. The first step of the procedure finds the relative maxima, i.e. crystals that have detected an energy higher than any of the adjacent cells and marks them progressively 1, 2, …, j, …, n. This means that, since the list is written and examined in descending energy order, the absolute maximum will be the seed of cluster number 1 and will carry a *crystal.clclu* $= 1$, the second highest relative maximum will be the seed of cluster number 2 and will carry a *crystal.clclu* $= 2$ and so on. At this stage we will have identified as many clusters as there are relative maxima. Each will contain a single member and will have *crystal.fever* equal to the energy deposited in the single member. In the following steps each established cluster j will be examined in descending E_{cl} order. Any of its already assigned members will be allowed to infect its neighbours by setting their *crystal.clclu* $= j$ if the previous *crystal.clclu* $= 0$. In this way when a crystal, that is not yet assigned to a cluster, is neighbour to more crystals already assigned to different clusters, it will be attributed to the cluster that has the highest cumulative E_{cl} energy. The procedure is iterated until there are no more neighbours to any of the n clusters to infect. At any step the list of crystals is re-ordered and examined by descending E_{cl} so that the cluster with higher cumulative energy has priority in infecting new adjacent cells. Up to this stage our method is equivalent to that of Breton et al. [7] and we will have grouped all active crystals around the n relative maxima. The main advantage of this procedure with respect to the contiguity method, is the ability of separating merging showers that deposit energy in a set of contiguous crystals but with distinguishable maxima. It is however evident that, without further action, on the crystal threshold E_{th} or on the procedure, spills and satellites cannot be spotted while distorted showers, with a clear relative maximum and a second weak relative maximum, due to statistical fluctuations in deposited energy more than to a merging second shower, are misinterpreted. We have added two more steps: one for distinguishing distorted clusters from merged showers, and one for identifying and incorporating energy spills.

For distorted clusters we proceed as follows. The clusters that are initiated by different relative maxima but are contiguous to some other cluster have the parameter *crystal.cnt* set to 1 (normal value 0). The list of clusters having *crystal.cnt* = 1 is examined in descending E_{cl} order. Suppose we are considering a cluster having a value $E_{\mathrm{cl}}(j_1)$ that is contiguous to a second cluster having a value $E_{\mathrm{cl}}(j_2) < E_{\mathrm{cl}}(j_1)$. We introduce a parameter R_{cnt} and we incorporate clusters j_1 and j_2 if $E_{\mathrm{cl}}(j_2)/E_{\mathrm{cl}}(j_1) < R_{\mathrm{cnt}}$. If not cluster j_2 will maintain its identity throughout. By doing this we will find less clusters than relative maxima. For eliminating spills we have to take into account clusters that, while being not contiguous, are relatively close in the θ–ϕ plane and of very different energy. For each cluster j including $k(j)$ crystals we compute the centroid co-ordinates $\theta_{\mathrm{C}}(j)$ and $\phi_{\mathrm{C}}(j)$ as follows:

$$\vartheta_{\mathrm{C}}(j) = \left\{ \frac{\sum_{l=1}^{k(j)} \vartheta_{\mathrm{cry}}(l) E_{\mathrm{cry}}(l)/V_{\mathrm{cry}}(l)}{\sum_{l=1}^{k(j)} E_{\mathrm{cry}}(l)/V_{\mathrm{cry}}(l)} - 25 \right\} \frac{1}{130}$$

$$\phi_{\mathrm{C}}(j) = \left\{ \frac{\sum_{l=1}^{k(j)} \phi_{\mathrm{cry}}(l) E_{\mathrm{cry}}(l)/V_{\mathrm{cry}}(l)}{\sum_{l=1}^{k(j)} E_{\mathrm{cry}}(l)/V_{\mathrm{cry}}(l)} \right\} \frac{1}{360}$$

where E_{cry} is the crystal energy, V_{cry} is the crystal volume and the centroid co-ordinates are scaled from 0 to 1. The Euclidean distance between two clusters is evaluated as $D(j_1, j_2) = \sqrt{(\vartheta_{\mathrm{C}}(j_1) - \vartheta_{\mathrm{C}}(j_2))^2 + (\phi_{\mathrm{C}}(j_1) - \phi_{\mathrm{C}}(j_2))^2}$. We introduce a further parameter $T_{\mathrm{DE\_}}$. Two cluster j_1 and j_2 will be considered as one if the dimensionless quantity

$$\frac{E_{\mathrm{cl}}(j_1) - E_{\mathrm{cl}}(j_2)}{E_{\mathrm{cl}}(j_1) + E_{\mathrm{cl}}(j_2)} \cdot \frac{1}{D(j_1, j_2)} > T_{\mathrm{DE}}.$$

By doing this we will find again less clusters than relative maxima

4.3. Deterministic annealing

While the previous two methods are associative methods in which decisions are taken at any step, the deterministic annealing paradigm performs the minimisation of an appropriate function and gives, at any step, an evaluation of the membership to clusters. Having m elements to be grouped in n clusters we can define a partition matrix $U = [u_{ji}]$ that gives, for element i $(1 \le i \le m)$ the association probability to cluster j $(1 \le j \le n)$. u_{ji} can assume any value between 0 and 1 with $\sum_{j=1}^{n} u_{ji} = 1$ (probabilistic constraint). In associative algorithms the association probability will assume either 0 or 1 value. In deterministic annealing the minimisation will proceed until a wanted condition on the entire partition matrix will be met. We must define a distortion function $E_j(\boldsymbol{x}_i)$ usually assumed as the squared Euclidean distance from the centroid of cluster j: $E_j(\boldsymbol{x}_i) = \|\boldsymbol{x}_i - \boldsymbol{y}_j\|^2$ and an expectation of the global error function $\langle E \rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} u_{ji} E_j(\boldsymbol{x}_i)$.

Following the maximum entropy approach by Jaynes [9] the function to maximise in the deterministic annealing approach is the Shannon entropy: $H(u_{11}, \dots, u_{1m}, \dots, u_{n1}, u_{nm}) = -K \sum_{i=1}^{m} \sum_{j=1}^{n} u_{ji} \ln u_{ji}$, where K is a constant. The association probabilities which maximise the entropy [10] are Gibbs distributions: $u_{ji} = \mathrm{e}^{-\beta E_j(x_i)}/Z_i$ where $Z_i = \sum_{l=1}^{n} \mathrm{e}^{-\beta E_l(x_i)}$ is a normalisation factor named partition function. We point out that those probabilities can be interpreted as memberships of fuzzy sets [11]. From a statistical mechanics point of view, the Lagrange multiplier β is interpreted as the inverse of temperature T $(\beta = 1/T)$. When β increases, the associations of elements to clusters become crisper and β can be interpreted as a control parameter of fuzziness. The limit cases are

- for $\beta \to 0^+$ we have $u_{ji} = 1/n$ for all j, i, i.e. each element is equally associated with each cluster,
- for $\beta \to +\infty$ we have $u_{ji} = 1$ if element i is associated to cluster j and zero elsewhere (hard limit).

Let us define the effective error (also named the free energy, by analogy to statistical mechanics) $F = -(1/\beta)\ln(\Pi_i Z_i)$. One can demonstrate that $F = -(1/\beta)H_{\max} + \langle E \rangle$ and then $\lim_{\beta \to \infty} F = \langle E \rangle$ [8]. This limit allows us to find the solution of the constrained minimisation of $\langle E \rangle$ by performing a so-called deterministic annealing on F. We start by minimising F at high T, for which there is a unique solution $u_{ji} = 1/n$ and then reduce T until the hard limit is reached. In practice we perform for each value of T a minimisation of F with respect to the set of clusters Y by iterating the following formulas

$$y_j = \frac{\sum_{i=1}^{m} u_{ji}\, \boldsymbol{x}_i}{\sum_{i=1}^{m} u_{ji}}, \quad u_{ji} = \frac{\mathrm{e}^{-\beta E_j(x_i)}}{\sum_{l=1}^{n} \mathrm{e}^{-\beta E_j(x_i)}}, \quad \forall j, i.$$

The procedure stops when any of the m elements has at least a membership larger than a wanted value M_{set} to one of the n clusters. At high temperature all centres collapse to a unique point and then, during annealing, natural clusters differentiate. Moreover under general conditions, every cluster will split at the critical temperature corresponding to its variance [10] giving rise to a natural hierarchical clustering.

5. Comparison of clustering algorithms

5.1. Adjustment of cellular automata parameters

Secondary particles, overlapping maps, split maps, distorted maps, are all sources of cluster counting errors that reduce the efficiency of reconstruction of the reaction final state, being the relative importance of each potential error source strongly dependent on the reaction channel and set threshold. The contiguity method cannot cope with split maps and overlapping maps, and is insensitive to distorted maps. The cellular automata technique, as we have modified it, can have the parameter T_{DE} adjusted for efficient recover of split maps, while the parameter R_{cnt} can be tuned to reach the best possible compromise between two opposite actions: the identification of merging maps and the identification of distorted maps. The deterministic annealing does not work simply on the topology of the events but its parameters should be tuned to avoid forcing the minimisation to unrealistic solutions, driven by the paucity of elements (crystals) on which the method will work in this particular application.

In Fig. 5a we have plotted the distribution of cluster energy ratios, equivalent to our R_{cnt} parameter for the case of two clusters identified in a single-particle map with two relative maxima (full curve) and for the case of two real merging clusters (dashed curve). It is evident that an appropriate selection of R_{cnt} should highly reduce the unwanted effect. We have used similar plots to set the most appropriate R_{cnt} value reported in Table 1 for the various reaction channels and thresholds. In Fig. 5b we show the distribution of clusters a-dimensional distances, equivalent to our T_{DE} parameter, in three different cases: primary particle from nearest primary particle (full curve), secondary particle from generating primary particle (dash–dot curve) and spills from main sub-map (dash curve).

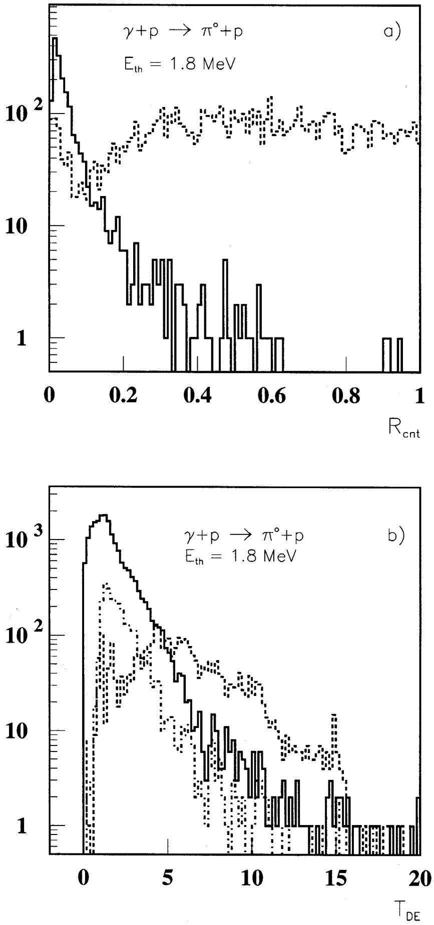

Fig. 5. Part a: the distribution of energy ratio between two clusters deriving from a wrong counting on single-particle maps showing two relative maxima (full curve) and from two merged particle maps (dash curve). Part b: the distribution of dimensionless distance (see text) between a primary particle and the closest primary (full curve), a secondary particle and its primary (dash-dot curve), a spill and the main sub-map of the same map (dash curve).

Table 1
Parameters used in cellular automata algorithm for the analysis of clusters in different reaction simulations

E_{th} (MeV)	R_{cnt} all reaction channels	R_{cnt} $\gamma + p \rightarrow \pi^0 p \rightarrow 2\gamma p$	R_{cnt} $\gamma + p \rightarrow \eta p \rightarrow 4\gamma p$	R_{cnt} $\gamma + p \rightarrow \eta p \rightarrow 6\gamma p$	T_{DE} all reaction channels
1.8	0.3	0.16	0.20	0.21	8.0
3.0	0.3	0.19	0.20	0.23	8.0
5.0	0.3	0.21	0.20	0.23	8.0
7.5	0.3	0.23	0.25	0.25	7.0
10.0	0.4	0.28	0.30	0.25	7.0
15.0	0.4	0.30	0.30	0.25	6.0

In this case the efficacy of a cut above T_{DE} is not as high as for R_{cnt}. We have used similar plots to set for the various reaction channels and thresholds the most appropriate T_{DE} value also reported in Table 1. The efficiency of identification of maps with more relative maxima by the R_{cnt} value is plotted as a full curve in Fig. 6a–d for all the considered cases and thresholds. The percentage of real merging maps that is unfortunately rejected by this selection is also shown in Fig. 6a–d as a dotted curve. We observe that, the parameter R_{cnt} listed in Table 1, assures a marginal constant rejection of real merging maps and a rather high efficiency in the identification of single maps with two relative maxima. Although the latter decreases with increasing threshold, the rejection of real merging maps can be accepted because of the relative importance of the effect.

The efficiency of identification of energy spills and their incorporation in the main map is plotted as a dash curve in Fig. 7a–d for all the considered cases and thresholds, using the T_{DE} parameter listed in Table 1. The percentage of reabsorbed spills is around 20–30% while the percentage of secondary particles reabsorbed by their primary (dot curve) and the percentage of primary particles unfortunately merged with the nearest primary (full curve) remains below 0.5% in almost all instances.

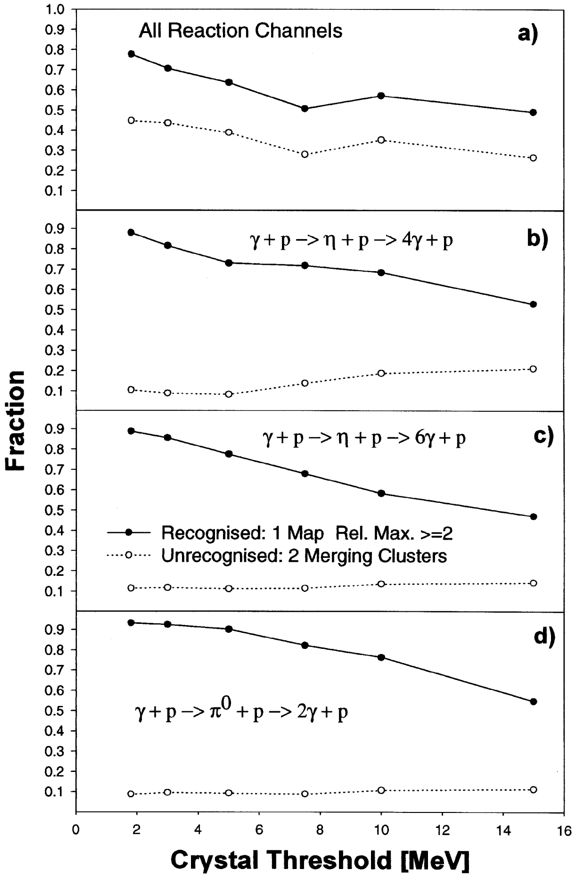

Fig. 6. The fraction of recognised single-particle maps with two maxima (full curve, full dots) and of unrecognised merging maps (dot curve, open circles) as a function of the energy threshold set for the 480 BGO crystals and for different simulated reaction conditions (parts a, b, c, d). The values are obtained by discriminating on the R_{cnt} parameter according to the values of Table 1.

5.2. Considerations on the deterministic annealing

We have investigated the deterministic annealing method for the same reactions outlined above, in order to determine the parameters most

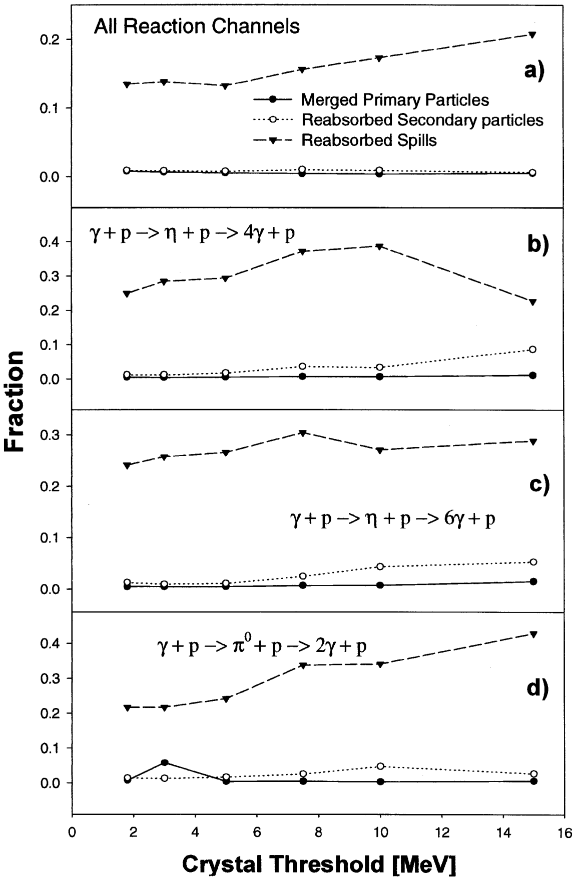

Fig. 7. The fraction of reabsorbed spills, of reabsorbed secondary particles and of mistakenly merged distinct primary particles as a function of the energy threshold set for the 480 BGO crystals and for different simulated reaction conditions (parts a, b, c, d). The values are obtained by discriminating on the T_{DE} parameter according to the values of Table 1.

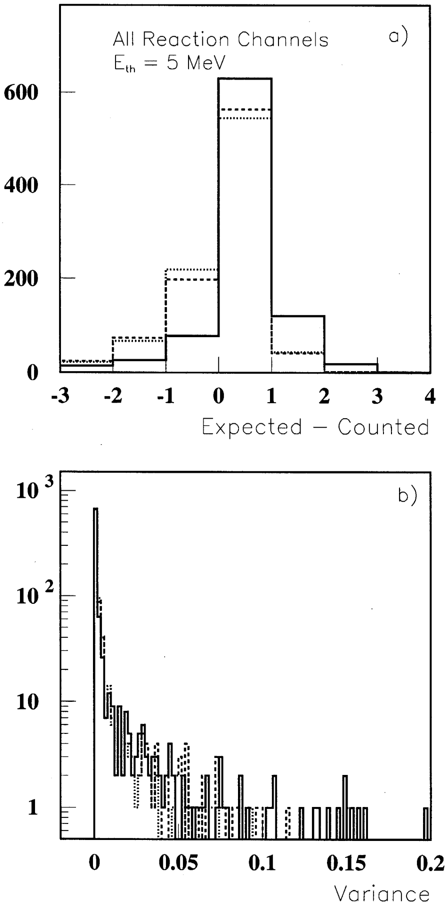

Fig. 8. The difference of expected to counted clusters (part a) and the energy variance (part b) obtained on 1 K events with the deterministic annealing (full curve), the cellular automata (dash curve) and the contiguity (dot curve) methods.

appropriate to a precise reconstruction of the reaction final state. As seen in Fig. 8a, we have found, on low statistics comparisons, that this method (full curve) is more efficient than the contiguity (dot curve) or the cellular automata (dash curve) in giving the correct number of clusters when the dominant channels are few body reactions, essentially single pion photoproduction. As regards the precision of energy reconstruction, the results are equivalent to the other two methods as seen in Fig. 8b. Unfortunately, the determination of parameters in deterministic annealing is strongly affected by the reaction channel under consideration, while for the other two methods the parameters show small variations as given in Table 1. Furthermore, the number of steps required for minimisation can be of the order of a few thousand, and the analysis time drastically increases with respect to the other methods. Since this would be a major problem in the analysis of experimental data for our set-up, which produces data blocks of the order of 1 million events, we have preferred to investigate thoroughly the two simpler methods, while being aware of the potential, high-efficiency applications of the deterministic annealing approach.

6. Results and discussion

The efficiencies obtained by the contiguity and cellular automata clustering methods, have been evaluated at high statistics as a function of crystal threshold, for the different reaction channels already described and, for the cellular automata method, with the parameter set of Table 1. In general terms we observe in Fig. 9a–d that the cellular automata gives an efficiency higher that the contiguity method, due to the capability of accounting for spurious effects. This is of course more evident at low thresholds, where both spills and secondary particles are present in higher percentages. The difference becomes marginal at high thresholds where the particle maps are inherently unstructured in the most general case. However, in the three selected photoproduction channels, the capability of the cellular automata to detect merging clusters is clearly put in evidence by the efficiency curves and makes the cellular automata preferable. The observed efficiencies go from 40–50% at low E_{th} up to 60–90% at high E_{th} for the contiguity method and from 50–60% at low E_{th} up to 70–90% at high E_{th} for the cellular automata method.

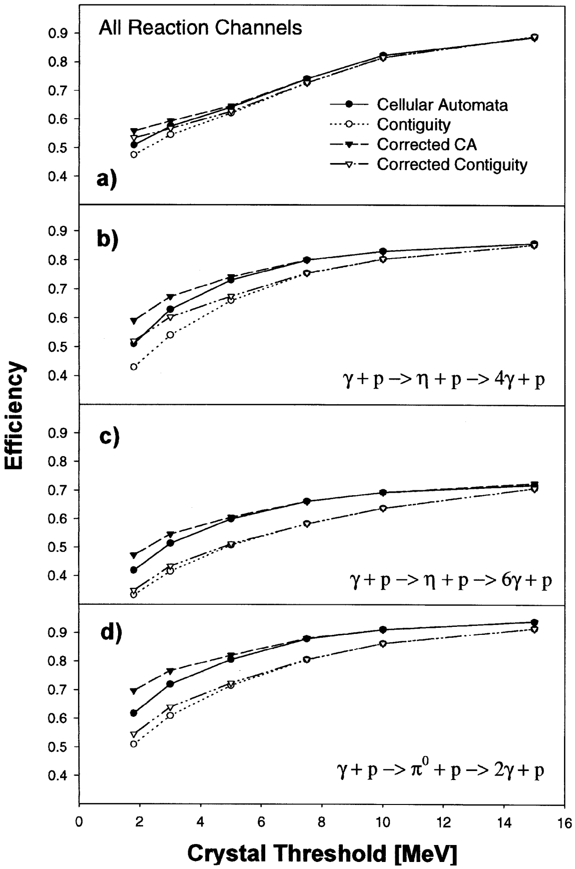

Fig. 9. The efficiency of cluster identification as a function of the energy threshold set for the 480 BGO crystals and for different simulated reaction conditions (parts a, b, c, d). The values obtained with the cellular automata and the contiguity methods are shown together with the values corrected on the basis of the percent accumulated energy.

We remind that our clusters are ordered by descending collected energy. We have extensively observed that the last identified cluster, often carries a very small portion ($< 1\%$) of the total collected energy. This could well be due to energy spills that are not accounted by the contiguity method and are corrected only to 20–40% by the T_{DE} parameter in the cellular automata method. Starting from this consideration we have set a new cluster counting rule, that takes, from the n clusters identified by the two methods only the $k < n$ that add up to more than 99% of the collected energy. As seen in Fig. 9a–d this further selection increases the efficiency by an extra 5–10% for both methods, mostly at low E_{th}, while at higher threshold this further selection is ineffective since we do not expect spills at such high thresholds.

The cellular automata method not only gives a better cluster number estimate but produces clusters whose energies are more close to the energy included in the particle separate map. This is shown in Fig. 10a where the distribution of the energy variance: $V = (1/E_{\mathrm{BGO}})\sum_{i=1}^{M}(E_{\mathrm{part}}(i) - E_{\mathrm{cl}}(i))^2$ has been plotted for the two methods. E_{BGO} indicates the energy collected for the event on the calorimeter, and M is the maximum between the number of particles and the number of counted clusters. The variance is lower when the exact number of clusters is predicted (Fig. 10a) and increases, equally for the two methods, when the cluster number is wrong (Fig. 10b). The variance, as a function of threshold, for the two methods, is plotted in Fig. 11.

A last consideration was made on the background produced by the different methods for specific reaction channels. We have considered in a set

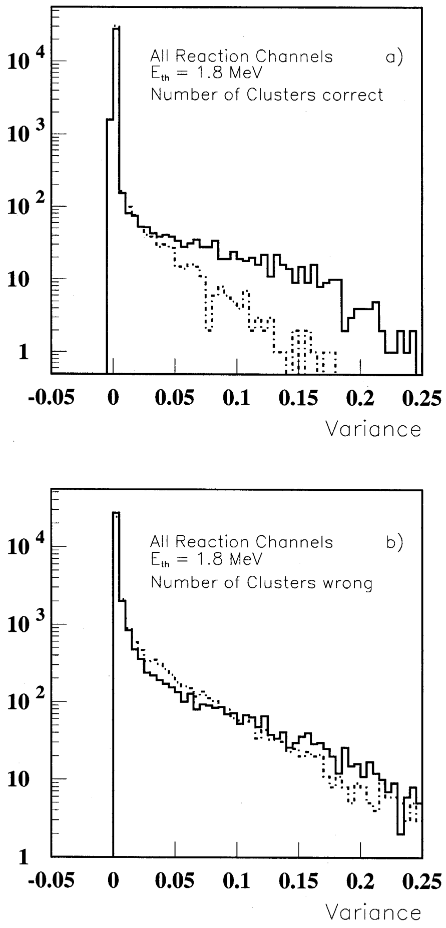

Fig. 10. Part a: the energy variance obtained by the contiguity method (full curve) and by the cellular automata method (dash-dot curve) when each method gives the correct number of clusters. Part a: the same when each method gives the wrong number of clusters.

of 10 000 simulated events, three cases and reported the results in Fig. 12:

(a) π^0 photoproduction with two particles detected in the calorimeter but accounted as one cluster. This error reduces the efficiency of reconstruction of the reaction channel. Fig. 12 reports (full dots, full line) the ratio between the number of events (out of 10 000) in which two particles are mistakenly counted as one by the contiguity method and the similar number found by the cellular automata method.

(b) π^0 photoproduction with three particles detected in the calorimeter but accounted as two clusters. This error reduces as well the efficiency of reconstruction of the reaction channel. Fig. 12 reports (open circles, dot line) the ratio between the number of events (out of 10 000) in which three particles are mistakenly counted as two by the contiguity method and the similar number found by the cellular automata method.

(c) η decay in $3\pi^0$ with six particles detected in the calorimeter, accounted as four clusters. This error is one of the major sources of background for the rare η decay mode $\eta \rightarrow \pi^0\gamma\gamma$. In this case we

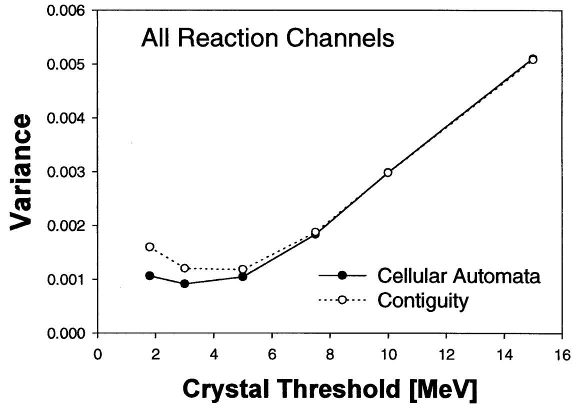

Fig. 11. The energy variance as a function of crystal energy threshold for the cellular automata and the contiguity methods.

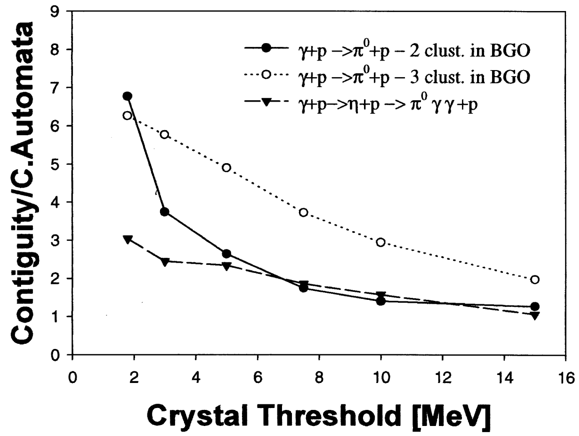

Fig. 12. Evaluation of some background effects in selected reaction channels as a function of crystal energy threshold. The ratio of values (see text for details) obtained by the contiguity method to the values obtained by the cellular automata method is plotted.

have computed for each of the two methods a sort of background to signal ratio (BS) by dividing the number of events (out of 10 000) in which the reaction $\eta \to 3\pi^0$ gives six particles in the calorimeter that are counted as four, and the number of events (out of 10 000) in which the reaction $\eta \to \pi^0\gamma\gamma$ gives four particles in the calorimeter that are counted exactly. Fig. 12 reports (full triangles, dash line) the ratio between the BS values obtained by the two methods.

It is evident that in any instance the cellular automata gives lower background effects, in particular in the examined rare η decay the background to signal ratio is three times smaller than for contiguity at low threshold and 60% smaller at high threshold.

7. Conclusions

We have applied different clustering algorithms to the analysis of the response of a segmented electromagnetic calorimeter, in use at the GRAAL-ESRF facility. With a complete, realistic simulation of the calorimeter response to photoreactions at intermediate energy and the addition of a few corrections to a standard cellular automata technique we have been capable of implementing an analysis code that offers advantages with respect to a simpler contiguity method and allows higher reconstruction efficiencies as well as reduced backgrounds in most reaction channels.

References

[1] GEANT Reference Manual, CERN Program Library, Long Write-up W5013 V3.21.

[2] FLUKA Users Guide, Technical Report TIS-RP-ISO, CERN, 1987, 1990.

[3] B. Girolami, F. Ghio, M. Castoldi, A. Zucchiatti, P. Levi Sandri, P. Rossi, V. Bellini, M. Capogni, L. Casano, L. Ciciani, A. D'Azeglio, A. D'Angelo, D. Moricciani, L. Nicoletti, C. Schaerf, R. Di Salvo, G. Gervino, Proc. VI Int. Conf. on Calorim. in High Ener. Phys. - Frascati Phys. Ser. VI, 1996, p. 727.

[4] D. Babusci, V. Bellini, M. Capogni, L. Casano, A. D'Angelo, F. Ghio, B. Girolami, L. Hu, D. Moricciani, C. Schaerf, Riv. Nuovo Cimento. 19 (1996) 5.

[5] P. Corvisiero, L. Mazzaschi, M. Ripani, M. Anghinolfi, V.I. Mokeev, G. Ricco, M. Taiuti, A. Zucchiatti, Nucl. Instr. and Meth. A 346 (1994) 433.

[6] P. Levi Sandri, F. Ghio, D. Moricciani, M. Breuer, M. Rigney, J.P. Didelez, Ch. Djalali, M. Anghinolfi, N. Bianchi, M. Capogni, L. Casano, P. Corvisiero, A. D'Angelo, E. DeSanctis, R. DiSalvo, G. Gervino, B. Girolami, L. Hu, V. Muccifora, E. Polli, A.R. Reolon, G. Ricco, M. Ripani, P. Rossi, M. Sanzone, C. Schaerf, M. Taiuti, A. Zucchiatti, Nucl. Instr. and Meth. A 370 (1996) 396.

[7] V. Breton, H. Fonvieille, P. Grenier, C. Guicheney, J. Jousset, Y. Roblin, F. Tamin, Nucl. Instr. and Meth. A 362 (1995) 478.

[8] K. Rose, E. Gurewitz, G. Fox, Pattern Recognition Letters 11 (1990) 589.

[9] E.T. Jaynes, Phys. Rev. Lett. 106 (1957) 620.

[10] K. Rose, E. Gurewitz, G. Fox, Phys. Rev. Lett. 65 (1990) 945.

[11] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum press, New York, 1981.