



**HAL**  
open science

## Development and implementation in the Monte Carlo code PENELOPE of a new virtual source model for radiotherapy photon beams and portal image calculation

Isabelle Chabert, Eric Barat, Thomas Dautremer, Thierry Montagu, Mathieu Agelou, S Croc de Suray, Juan Garcia-Hernandez, S Gempp, M Benkreira, Loïc de Carlan, et al.

### ► To cite this version:

Isabelle Chabert, Eric Barat, Thomas Dautremer, Thierry Montagu, Mathieu Agelou, et al.. Development and implementation in the Monte Carlo code PENELOPE of a new virtual source model for radiotherapy photon beams and portal image calculation. *Physics in Medicine and Biology*, 2016, 61 (14), pp.5215 - 5252. 10.1088/0031-9155/61/14/5215 . hal-01867678

**HAL Id: hal-01867678**

**<https://hal.science/hal-01867678v1>**

Submitted on 30 Sep 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Development and implementation in the Monte Carlo code PENELOPE of a new virtual source model for radiotherapy photon beams and portal image calculation

I Chabert<sup>1</sup>, E Barat<sup>1</sup>, T Dautremet<sup>1</sup>, T Montagu<sup>1</sup>, M Agelou<sup>1</sup>, A Croc de Suray<sup>1</sup>, J.C Garcia-Hernandez<sup>1</sup>, S Gempp<sup>2</sup>, M Benkreira<sup>2</sup>, L de Carlan<sup>1</sup>, and D Lazaro<sup>1</sup>

<sup>1</sup>CEA, LIST, F-91191 Gif-sur-Yvette, France

<sup>2</sup>AP-HM, Hôpital Nord, Chemin des Bourrely, 13015 Marseille, France

E-mail address: [ISABELLE.CHABERT@gustaveroussy.fr](mailto:ISABELLE.CHABERT@gustaveroussy.fr), [Eric.BARAT@cea.fr](mailto:Eric.BARAT@cea.fr); [Thomas.DAUTREMER@cea.fr](mailto:Thomas.DAUTREMER@cea.fr); [Thierry.MONTAGU@cea.fr](mailto:Thierry.MONTAGU@cea.fr); [mathieu.agelou@cea.fr](mailto:mathieu.agelou@cea.fr); [aurelien.croc@cea.fr](mailto:aurelien.croc@cea.fr); [juan-carlos.garcia-hernandez@cea.fr](mailto:juan-carlos.garcia-hernandez@cea.fr); [Stephanie.RAUCOULES@ap-hm.fr](mailto:Stephanie.RAUCOULES@ap-hm.fr); [Mohamed.BENKREIRA@ap-hm.fr](mailto:Mohamed.BENKREIRA@ap-hm.fr); [loic.lenoir-de-carlan@cea.fr](mailto:loic.lenoir-de-carlan@cea.fr); [delphine.lazaro@cea.fr](mailto:delphine.lazaro@cea.fr)

## Abstract

This work aims at developing a generic virtual source model (VSM) preserving all existing correlations between variables stored in a Monte Carlo pre-computed phase space (PS) file, for dose calculation and high-resolution portal image prediction. The reference PS file was calculated using the PENELOPE code, after the flattening filter (FF) of an Elekta Synergy 6 MV photon beam. Each particle was represented in a mobile coordinate system by its radial position ( $r_s$ ) in the PS plane, its energy ( $E$ ), and its polar and azimuthal angles ( $\varphi_d$  and  $\theta_d$ ), describing the particle deviation compared to its initial direction after bremsstrahlung, and the deviation orientation. Three sub-sources were created by sorting out particles according to their last interaction location (target, primary collimator or FF). For each sub-source, 4D correlated-histograms were built by storing  $E$ ,  $r_s$ ,  $\varphi_d$  and  $\theta_d$  values. Five different adaptive binning schemes were studied to construct 4D histograms of the VSMs, to ensure histogram efficient handling as well as an accurate reproduction of  $E$ ,  $r_s$ ,  $\varphi_d$  and  $\theta_d$  distribution details. The five resulting VSMs were then implemented in PENELOPE. Their accuracy was first assessed in the PS plane, by comparing  $E$ ,  $r_s$ ,  $\varphi_d$  and  $\theta_d$  distributions with those obtained from the reference PS file. Second, dose distributions computed in water, using the VSMs and the reference PS file located below the FF, and also after collimation in both water and heterogeneous phantom, were compared using a 1.5% / 0 mm and a 2% / 0 mm global gamma index, respectively. Finally, portal images were calculated without and with phantoms in the beam. The model was then evaluated using a 1% / 0 mm global gamma index. Performances of a mono-source VSM were also investigated and led to excellent results in the case of an adaptive binning.

**PACS code:** see <http://www.aip.org/pacs>: 87.10.Rt (Monte Carlo simulations), 87.55.Qr (Quality assurance in radiotherapy), 87.56.Fc (Quality assurance equipment), 02.50.-r (Probability theory, stochastic processes, and statistics), 02.60.Ed (Interpolation; curve fitting).

**Submitted to:** Physics in Medicine and Biology

## 1. Introduction

Advanced techniques used in radiotherapy for cancer treatment, such as Intensity-Modulated Radiation Therapy (IMRT), require extensive verification procedures to ensure the correct dose delivery. Electronic Portal Imaging Devices (EPIDs) are widely used for quality assurance in radiotherapy, and also for dosimetric verifications. For this latter application, the images obtained during the treatment session can be compared to a pre-calculated reference image in order to highlight dose delivery errors. The Monte Carlo (MC) method is widely used to perform accurate modeling of the radiotherapy dose delivery process, as well as for portal image prediction. However, this accuracy goes hand in hand with important computation time incompatible with clinical routine. The tracking of particles in the linac head takes most of the calculation time. In order to reduce it, a common practice consists in dividing the MC modeling of the linac head in two steps. This two-step calculation takes advantage of the fact that the tracking of particles in the upper part of the linac head, including static components whatever the delivered treatment (target, primary collimator, and flattening filter), could be performed only once. The physical characteristics of all exiting particles (energy, position and direction) are then stored at this point in a beam phase space file (PSF). All the information needed is then available in this PSF to continue particle tracking in the lower part of the linac head, and dose calculations in the patient geometry.

However, three main drawbacks are associated to the use of PSFs. First, their use requires huge storage capacities since a single PSF, describing a given radiotherapy beam, could be as large as hundreds of gigabytes. Another problem concerns the limited number of particles stored in the PSF, leading to potential systematic errors in dose calculations, and even biases if the PSF is replayed too often (Sempau et al., 2001). Finally, the mode of representation of the physical characteristics of the linac beam in a PSF is not flexible enough to account for variations between individual linacs of the same model.

Multiple-source models were introduced as an attractive alternative solution to overcome part or all these limitations. They consist in characterizing the beam using several sub-sources, each of them representing a major component of the linac head, and in grouping particles having the same physical characteristics. These multiple-source models, also called virtual source models (VSMs), can be constructed either from a full MC simulation of the linac, with or without enhancement from measured data (Ma et al., 1997; Schach von Wittenau et al., 1999; Deng et al., 2000; Fix et al., 2000, 2001a, 2001b, 2004), or derived exclusively from measured data (Ahnesjö et al., 2005; Cashmore et al., 2012; Faddegon and Blevis, 2000; Fippel et al., 2003; Janssen et al., 2001; Sikora et al., 2007; Tillikainen and Siljamäki, 2008). Schach von Wittenau (Schach von Wittenau et al., 1999) proposed in their pioneering work a compact but accurate VSM representation, in which correlated histograms represent the distributions in energy, direction and position, for each sub-source. The advantage of this approach is to retain the correlations existing among the variables stored in the PSF, hence preserving the level of details brought by MC simulations, but in a more compressed format.

The major challenge facing histogram-based VSMs is that the analytical or parameterized functions used in histogram-based VSMs must preserve as much as possible information stored in the PSF about correlations between energy, position and angular distributions of the particles, with respect to the required level of accuracy for the dose calculations. Indeed, the smaller the pixel or voxel size, the more accurate should be the VSM in the description of the physics. If one wants to retain all correlations existing between particle characteristics, multi-dimensional histograms must be generated, and their handling in the VSM could become complex and sensitive. In particular, the number of bins they contain is a tricky parameter to set in the source model. It is obvious that the higher the number of bins, the higher the accuracy with which a histogram will represent a given distribution. Nevertheless, this number of bins should be limited and controlled, mainly because the statistical uncertainty of each bin increases with the number of bins, and the associated risk is to construct very noisy correlated histograms. The distribution of the bins for each variable, or binning scheme, represents a major problem in the development of a VSM, and should be carefully chosen to balance a tradeoff between the size of the histograms, their statistical uncertainty and the required accuracy of the VSM. In order to tackle this problem, the VSMs previously proposed (Deng et al., 2000; Fix et al., 2004, 2001a, 2001b; Schach von Wittenau et al., 1999) reduced the dimensions of the problem by removing at least one of the correlations that seemed to be less important. This reduces the number of bins in the histograms, sometimes at the expense of accuracy, depending on the conditions of use of the VSM.

We chose to explore another way to construct a histogram-based VSM, with the goal to retain all the existing correlations between variables stored in the PSF. An efficient handling of the multi-dimensional histograms was made possible by using an adaptive binning to construct the correlated histograms. The PSF can be seen as a statistical sample of the probability density function describing the phase space distribution. The main goal when designing a beam source, either in the form of a PSF or a source model, is to obtain the most robust estimator for the phase space distribution. The binning of the histograms should hence be carefully chosen in order to reproduce as accurately as possible the details of the energy, position and angular distributions. Different adaptive binning schemes were tested, and the performances of the associated VSMs were evaluated and compared to those of a regular binning. The validation of the VSMs was first performed by comparing radial, angular, and energy distributions calculated in the phase space plane using the VSMs with those obtained from a reference PSF. 3D dose distributions were then computed in a water phantom located below the flattening filter, using the VSMs and the reference PSF, and also after collimation systems in the same water phantom and in a heterogeneous phantom. Associated portal images were finally calculated. The difficulties implied here are related to the extremely high resolution of the EPID, its pixel size being  $0.4 \times 0.4 \text{ mm}^2$ , coupled to its position far away from the source. In this configuration, the VSM may be pushed to the limits with respect to dosimetric accuracy. This paper also addresses the possibility of representing accurately the beam with a unique source and adapted binning.

## 2. Material and methods

### 2.1. Description of the virtual source model

The VSMs developed and studied in this work are histogram-based, and the construction of the correlated histograms was derived from the analysis of information stored in the PSF obtained from a full MC simulation of the linac treatment head. The VSM includes three sub-sources describing the target (the primary source), the primary collimator and the flattening filter (the scattered sources) using a set of correlated histograms. Particles stored in the PSF were then sorted out into these three sub-sources, according to the location of their last interaction. The intensity of each sub-source was given by the ratio of the number of particles belonging to the sub-source over the total number of particles. Charged particles were neglected since their contribution represents only 0.6% of the total number of particles stored in the PSF for a 6 MV photon beam.

Each particle of each sub-source was described by its energy  $E$ , its position in the phase space plane and its direction. Assuming the linac axial symmetry the particle position can be represented only by its radius  $r_s$ . The direction was not expressed in the PENELOPE coordinate system, referred in the following to as “linac coordinate system”, but in a mobile coordinate system related to the particle position and denoted as “the VSM coordinate system” in the following. The VSM coordinate system was defined by the line from the center of the target to the point where the particle crossed the phase space plane  $P$ , and by the normal to this line intercepting the linac axis in  $Q$  (cf figure 1). Polar and azimuthal angles ( $\varphi_d$ ,  $\theta_d$ ) were used in this system to represent the particle deviation with respect to a virtual direction corresponding to a bremsstrahlung event located at the center of the target  $O$ , and the orientation of this deviation (figure 1). The VSM coordinate system was chosen because it allows to better figure out the behavior of particles in the linac head. In particular, bremsstrahlung photons induce very small  $\varphi_d$  values, only due to their interaction location in the target. This leads to a neat discrimination of polar deviation distributions between bremsstrahlung and scattered photons. As explained later, this property will allow a choice of histogram binning matched to each source. Furthermore, this representation tends to make marginal distributions of  $\varphi_d$  and  $\theta_d$  rather informative. Finally, dependencies between variables  $r_s$ ,  $\varphi_d$  and  $\theta_d$  are not troubled by symmetries induced by the fixed reference system. The variable dependencies in the VSM coordinate system tend then to exhibit intrinsic properties of the linac design.

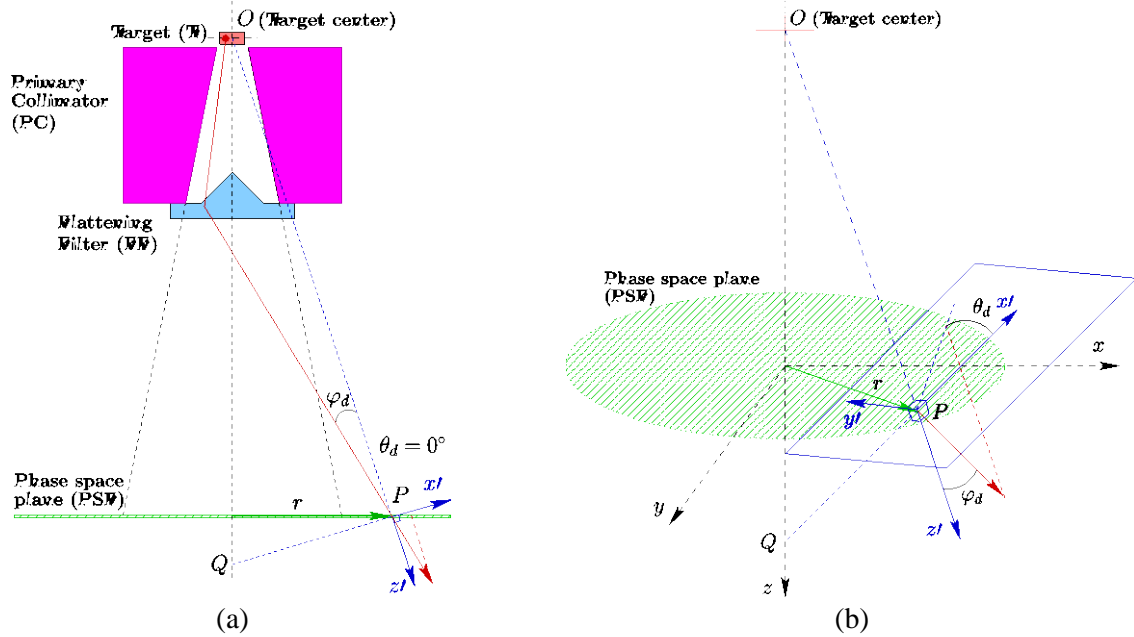


Figure 1. Description of the mobile coordinate system chosen to represent particles in the VSM. (a) A cross-section containing the particle and the linac axis ( $\theta_d = 0^\circ$ , here), showing a case where particle last interaction took place in the flattening filter. (b) A pure geometric perspective representation of the particle direction in the mobile coordinate system.

Each particle was hence described by the four abovementioned variables ( $r_s$ ,  $E$ ,  $\varphi_d$ ,  $\theta_d$ ) in the VSM coordinate system, and the phase-space function could be written as  $F(r_s, E, \varphi_d, \theta_d)$ .

In order to keep all the correlations between these four variables in the histogram-based model, it was necessary to describe each sub-source used to construct the VSM by a 4D-histogram. Hence, our VSM is based on a set of three 4D-histograms.

## 2.2. Creation of the VSM

### Generation of the PSF using the MC code PENELOPE.

The first step in the construction of the VSM consists in creating the PSF by full MC simulation of the linac head. The 6 MV photon mode delivered by a Synergy linac and its multi-leaf collimator MLCi2 (Elekta, Crawley, UK) and its associated EPID were then modelled using the MC simulation code PENELOPE. The treatment head is composed of a target (T), a primary collimator (PC), a flattening filter (FF), a monitor chamber, a mirror, the multi-leaf collimator (MLC) and its associated backup jaws, and the jaws. The geometry and materials of the different components were described based on information given by the manufacturer. The primary electron source impinging on the target was defined using a Gaussian-shaped energy distribution with a 5.95 MeV mean electron energy and a 0.5 MeV full width at half maximum. A uniform distribution of 1.5 mm radius was set for the spatial distribution of the incident electron beam. Electron and photon threshold energy cutoffs were set to 500 keV for electrons/positrons, and 10 keV for photons in PENELOPE. Parameters  $C_1$  and  $C_2$ , determining the mean free path and between hard elastic events, and the maximum average fractional energy loss in a single step, respectively, were both set to 0.05. The cutoff energy loss for inelastic collisions and bremsstrahlung emission were set to  $W_{cc} = W_{cr} = 10$  keV. DSMAX, representing the maximal range of a charged particle between two interaction is set to 0.001 cm in the target. No variance reduction techniques were used in the simulations.

Particles were transported through the treatment head until they crossed a cylindrical phase space plane of 20 cm radius, located at 16 cm from the electron source and just below the FF. All the information of each particle reaching this plane was stored in a PSF, and slightly differs from one MC code to another. In the PENELOPE code, each particle is described in the PSF by a binary record of 92 bytes storing fifteen variables:  $KPAR$ ,  $E$ ,  $X$ ,  $Y$ ,  $Z$ ,  $U$ ,  $V$ ,  $W$ , *weight*,  $ILB1$ ,  $ILB2$ ,  $ILB3$ ,  $ILB4$ ,  $ILB5$ ,  $NSHI$ , where  $KPAR$  represents the kind of particle,  $E$  its total energy,  $X$ ,  $Y$ ,  $Z$ , its Cartesian coordinates

in the phase space plane,  $U$ ,  $V$ ,  $W$  the cosines of the direction of movement,  $weight$  its statistical weight, and  $NSHI$  the incremental shower number. The five  $ILB$  variables enable the tracking of the particle in the geometry by giving information about its history. In particular, the  $ILB5$  variable was modified to enable the record of the body or volume where each particle underwent its last interaction.  $ILB5$  was encoded as  $3000*IBODYL + IBODYC$ , where  $IBODYC$  designed the body where the particle was created and  $IBODYL$  the body where its last interaction took place. Thanks to this modification, particles could be sorted out according to the location of their last interaction. It can be noticed that any MC simulation code enabling to retrieve the information about the location of the last interaction of a particle could be used to create this VSM. For instance, the  $LATCH$  variable could be used in the EGSnrc code for this purpose. A PSF corresponding to the storage of  $8.10^8$  photons in a 71 Go file was computed and analyzed to create the VSMs.

*2.2.1. Creation and storage of the correlated 4D-histograms.* The 4D-histograms were created by using an in-house program performing an automatic analysis of the information contained in the PSF. This program starts by sorting out the particles according to their type (photon, electron or positron). If the particle is a photon, the body of the last interaction is identified using the  $ILB5$  variable, and the particle was assigned to the corresponding sub-source (T, PC or FF). The particle energy  $E$  is then used to identify the corresponding energy bin, noted  $i(E)$ . The radial position of the particle in the phase space plane  $r_s$  is calculated using the  $X$  and  $Y$  particle coordinates and equation 1, and the corresponding radial position bin  $i(r_s^2)$  is identified. The  $r_s$  variable was first directly stored, but to take into account the fact that an infinitesimally small annular area is proportional to  $r.dr$ , we finally chose to store  $r_s^2$  instead of  $r_s$  to overcome this problem. The direction of the particle is expressed in the VSM coordinate system by calculating the values of  $\varphi_d$  and  $\theta_d$  angles, and the corresponding angle bins  $i(\varphi_d)$  and  $i(\theta_d)$  are identified. The calculation of the polar and azimuthal angles is detailed in appendix A.

$$r_s = \sqrt{X^2 + Y^2} \quad (1)$$

Once the four bins  $i(r_s^2)$ ,  $i(E)$ ,  $i(\varphi_d)$ ,  $i(\theta_d)$  identified for all particles, several distributions can be built for each source, representing piecewise constant probability densities. 1D-histograms represent marginal distributions describing individually the particle behaviour in terms of radial position, energy, polar or azimuthal angles. 2D-, 3D-, and 4D-histograms are joint probability density functions which represent the particle correlated behaviour with respect to two of the four variables, three of the four variables, and for all variables, respectively.

In order to prepare the handling of 4D-histograms in PENELOPE, each 4D-correlated histogram is processed so that it could be represented in the form of a 1D table. The position of each particle in this table, noted  $i_{4D}$ , is given by equation 2:

$$i_{4D} = i(r_s^2) + i(E).n(r_s^2) + i(\varphi_d).n(r_s^2).n(E) + i(\theta_d).n(r_s^2).n(E).n(\varphi_d) \quad (2)$$

Where  $n(r_s^2)$ ,  $n(E)$ ,  $n(\varphi_d)$  and  $n(\theta_d)$  refer to the number of bins for the radial, the energy, the polar angle and the azimuthal angle variables, respectively.  $i_{4D}$  hence represents the  $i^{th}$  bin in the 1D table.

Each 4D-histogram is stored in a separated text file containing in a first part general information, such as the sub-source intensity, the location of the phase space plane along the z-axis, the number of bins  $n(r_s^2)$ ,  $n(E)$ ,  $n(\varphi_d)$  and  $n(\theta_d)$  for each of the four variables, and the border values of each bin in the four dimensions. The values for each bin of the 4D-histogram are then stored. Let us notice that the sum of the sub-source intensities has to be equal to 1.

*2.2.2. Description of the studied binning schemes.* The goal of our study is to represent the probability density function describing the phase space using a VSM. The PSF calculated by the MC simulation constitutes one observation of this phase space function, from which we try to build the most robust and accurate estimator as possible, *i.e* the VSM. The quality of this estimator depends on several parameters. Among them, the choice of the binning for the 4D-histograms plays an important role since a trade-off has to be found between the dosimetric accuracy required within the

context of the application, good computing performances during the simulation runtime, and acceptable storage. The binning scheme drives directly the accuracy of the VSM since it must be flexible enough to adapt to the different nature of distribution-functions (energy, position and angular distributions) stored in 4D-correlated histograms, for each sub-source. Different binning schemes were thus developed and tested to construct the 4D-histograms. The size of the considered PSF also impacts directly the statistical uncertainty in the 4D-histograms. Although more compact in its representation, histogram-based VSMs have, however, intrinsic limitations in the size of the histograms, which in turn limits the number of bins. In fact, one should keep in mind that the 4D-histograms of the VSM would have to be loaded in the RAM memory during the initialization step. In order to respect the specifications of our computing cluster, we limited each 4D-histogram to  $64^4$  bins (64 bins for each of the four variables): one single 4D-histogram is stored on 100 Mo and requires about 400 Mo of RAM memory. For the three 4D-histograms included in our source model, 1.2 Go of random memory would be required. If the number of bins is increased to 128, the required storage for a single 4D-histogram would be about 2 Go, hence 6 Go for the three 4D-histograms and 24 Go of RAM memory. The user should therefore be aware that the histogram size cannot be larger than a limit which depends on the intrinsic limitations of its own computing machine.

Once the bin number in the histogram fixed, our goal was to know how to fix the width of the bins in order to reproduce the most accurately as possible the shapes of the distribution functions. This optimization was done for each of the VSM sub-sources. Four adaptive binning methods were tested and compared to a regular binning in order to determine the most suited binning. They were all based on the analysis of the particle distributions in a single dimension. The binning for each variable was chosen without taking into account the binning of other variables.

**Regular binning.** The first binning method, noted *Reg* in the following, consists in setting the bin width as constant. This constant was simply calculated as the maximum value of the variable divided by the number of bins. The bin size in the derived 4D-histogram is therefore exactly the same everywhere and depends only on the sub-source. The VSM constructed using the *Reg* binning is noted  $VSM^{Reg}$  in the following.

**Approximation of the Cumulative distribution.** The second method, noted *Cumul* in the following, uses the 1D-histograms obtained with very thin bins to calculate the Cumulated distribution and to fit it using a piecewise-linear function. In order to define each piece of this function, the deviation between the Cumulated distribution and the fitted estimation at each point was computed and compared to a deviation value set by the user. Each piece of the piecewise-linear function represents a bin, and the number of bins could be smaller than 64. The precision of this method depends on the number of pieces. The advantage of this method is the possibility to promote a finer binning in a direction when necessary. Table 1 sums up the characteristics of each piecewise-linear function used to fit the distributions of each sub-source for the VSM.

**Table 1.** Characteristics of the Cumulative distributions -  $N_{points}$  corresponds to the number of pieces used to approximate the original distribution.  $N$  is the number of points in the original distribution and  $d$  the step width. The maximum tolerated difference between the Cumulated distribution and its approximation is  $Err_{max}$ .

	Squared position		Energy		Polar angle		Azimuthal angle	
	$N_{points}$	$Err_{max}$	$N_{points}$	$Err_{max}$	$N_{points}$	$Err_{max}$	$N_{points}$	$Err_{max}$
Target	70	0,01 %	72	0,05 %	55	0,001 %	62	0,01 %
PC	61	0,04 %	68	0,1 %	55	0,05 %	71	0,01 %
FF	61	0,03 %	59	0,05 %	65	0,05 %	78	0,01 %
$N - d$	4000	0.1 cm	1600	5 keV	1200	0.1 deg	1800	0.1 deg

An example is given on figure 2 for the target energy distribution. The original cumulated distribution was described using 1600 points, but only 72 pieces were necessary to approximate it with a maximum error of 0.05%. The VSM constructed using the *Cumul* binning is noted  $VSM^{Cumul}$  in the following.

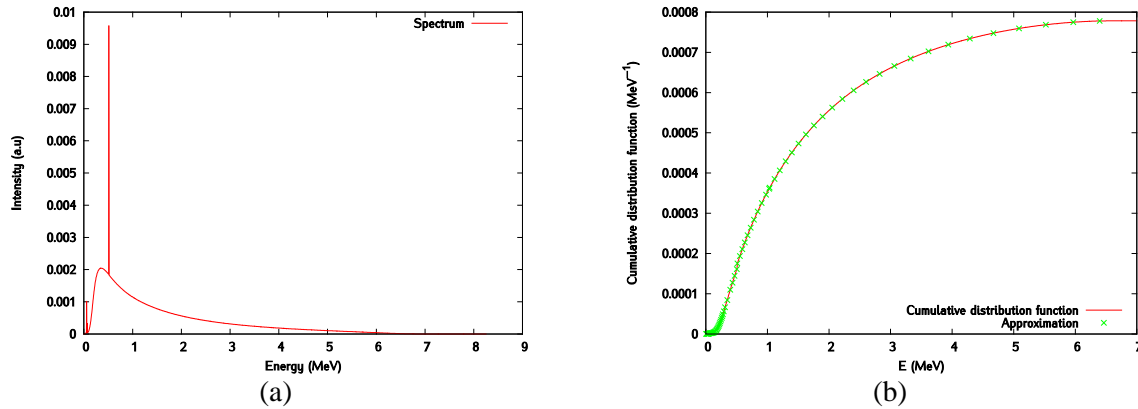


Figure 2. (a) Target spectrum (1600 points); (b) Cumulated spectrum (1600 points) and its approximation (72 points).

**Quantile method.** The Quantile method, noted *Quant*, induces an equal area between the bins in the 1D-distributions. The number of particles in each bin of the 1D-histogram is also strictly the same. The advantage of this method is the control of the statistical uncertainty in each bin. The VSM constructed using the *Quant* binning is noted  $VSM^{Quant}$  in the following.

**Likelihood maximization (ML) method.** The Likelihood Maximization method, noted *ML*, consists in working with irregular bins whose widths are inferred by the maximization of the likelihood. The detailed process used to determine this binning is explained in Appendix C. The VSM constructed using the *ML* binning is noted  $VSM^{ML}$  in the following.

**Penalized likelihood maximization method (MAP).** The Penalized Likelihood Maximization method, noted *MAP* (*for Maximum A Posteriori*), combines the *Quant* binning and the *ML* binning. The bin widths are inferred with respect to a penalized likelihood criterion. The chosen penalty is a prior encouraging bins of equal areas, modeled by a Dirichlet distribution. The detailed process used to determine this binning is explained in Appendix C. The VSM constructed using the *MAP* binning is noted  $VSM^{MAP}$  in the following.

**2.2.3. Implementation of the VSM in the PENELOPE code.** The implementation of the VSM into the PENELOPE code can be split into two parts. First, the initialization step consists in reading and processing the 4D-histograms in order to optimize the sampling. The second step consists in generating new particles, each of them being considered as a new shower. This is different from usual MC simulations where a primary electron is at the origin of a new shower. The number of simulated showers has no impact on the dose calculation, but the uncertainty calculation must be normalized, by applying a ratio defined as the number of emitted bremsstrahlung photons per primary electron. This ratio depends on the parameters of the primary electron beam (energy and spatial distributions) and on the variance reduction techniques used in the MC calculation. Its value in our study was 0.059.

**Histogram initialization.** The number of sub-sources is first identified and a test is performed to check that the sub-source intensities sum is actually equal to unity. 4D-histograms cannot be used in their initial form because the sampling process would be too slow. In order to speed up the sampling, we implemented the Walker's Alias Method improved by Peterson and Kronmal (Devroye, 1986, p. 107), which allows reorganizing the 4D-histograms in a more suited way. Indeed, this kind of methods implies a preprocessing time which increases linearly with the number of bins while the time devoted to sampling becomes independent of the number of bins.

**Use of the 4D-histograms.** Once the 4D-correlated histograms preprocessed, particles were generated using the sampling procedure illustrated in figure 3:



- Step 1 – The sub-source is chosen according to its intensity thanks to a uniformly sampled random number between 0 and 1.  $i(r_s^2)$ ,  $i(E)$ ,  $i(\varphi_d)$ ,  $i(\theta_d)$
- Step 2 - A bin  $i_{4D}$  is sampled from the 4D-correlated histogram of the selected sub-source, from which the radial position bin  $i(r_s^2)$ , the energy bin  $i(E)$ , the polar angle bin  $i(\varphi_d)$  and the azimuthal angle bin  $i(\theta_d)$  were obtained by successive Euclidean divisions.
- Step 3 - The particle coordinates are uniformly sampled in each of these bins:  $r_s^2$ ,  $E$ ,  $\varphi_d$ ,  $\theta_d$ . The sign of the azimuthal angle is randomly chosen.
- Step 4 – If needed, it is possible for the user to restrict the use of the VSM to a smaller radial value, noted R, than that initially set in the PSF plane ( $r = 20$  cm at  $z = 16$  cm). If  $r$  is smaller than the chosen R value, the particle is considered in the model. Otherwise, a simplified Russian roulette is applied to the particle. A random number is sampled uniformly between 0 and 1. The particle survival probability, noted  $P_{RR}$ , can be set by the user. If the random number is greater than  $P_{RR}$ , the particle survives and its weight is equal to  $1/(1-P_{RR})$ . If the random number is smaller than  $P_{RR}$ , the particle is killed by setting its weight to zero.
- Step 5 - These coordinates are expressed in the VSM coordinate system, and are converted into the linac coordinate system (see Appendix B for details).

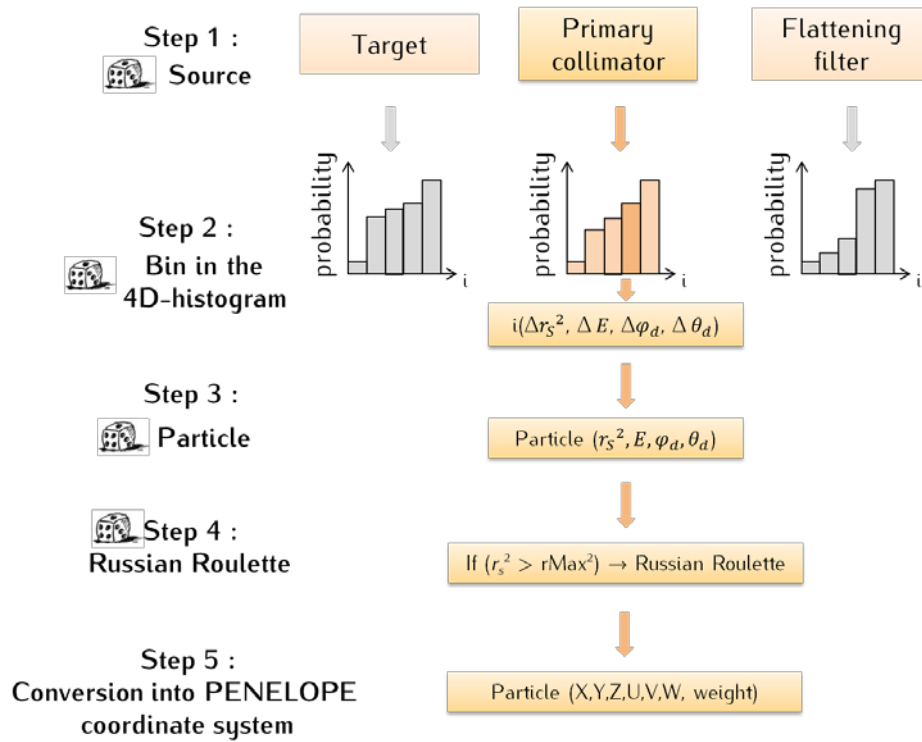


Figure 3. Process of creation of a new particle using the VSM. In this example, the chosen sub-source is the PC.

### 2.3. Assessment of the VSM performances

2.3.1. *Physical interpretation of correlated histograms and analysis of correlations.* A Hoeffding test of independence (Wilding and Mudholkar, 2008) was used to highlight the existing correlations between the variables  $r_s$ ,  $E$ ,  $\varphi_d$ ,  $\theta_d$ . This nonparametric rank test is well-suited here because it is consistent for any bivariate alternatives. The Hoeffding test estimates the difference H between the joint distribution function and the product of the marginal distributions of two variables  $v_1$  and  $v_2$ , as shown in equation 3:

$$H = \int (F_{12} - F_1 F_2)^2 dF_{12} \quad (3)$$

Where  $F_{12}$  is the joint distribution of two variables and  $F_i$ , their marginal distribution functions. In our case,  $F_{12}$  refers to the 2D-histogram of the variables  $v_1$  and  $v_2$ , and  $F_1$  and  $F_2$  to their 1D-histograms. A table containing p-values could be calculated: if the p-value is smaller than the chosen significance level (set to 0.01), the difference is significant,  $v_1$  and  $v_2$  can be said dependent. Pairwise tests were performed to show that two-by-two correlations exist between  $r_s$ ,  $E$ ,  $\varphi_d$  and  $\theta_d$ . A random sample of  $2^{13}$  particles for each sub-source was used.

The advantages of the Hoeffding test over other statistical tests, such as the Pearson Correlation Coefficients test or the Spearman test, are that correlations become visible when examining a very small portion of the PSF particles ( $2^{13}$ ). Moreover, the Hoeffding test enables to characterize any correlation, whatever its nature, whereas other tests make hypothesis about the nature of the sought out dependences (linear for the Pearson test, evolution of variations in the distributions for the Spearman test).

**2.3.2. Comparison of the binning schemes and dosimetric validation of the VSM.** The goal is twofold: first, we determined which binning schemes were able to reproduce the information stored in the PSF, and we assessed the dosimetric accuracy of the  $VSM^{Reg}$ ,  $VSM^{Cumul}$ ,  $VSM^{Quant}$ ,  $VSM^{ML}$  and  $VSM^{MAP}$ . Second, we validated the implementation and use of these new VSMs in the PENELOPE code. This validation procedure was performed following three steps. First, the radial, angular, and energy distributions in the linac coordinate system obtained using  $VSM^{Reg}$ ,  $VSM^{Cumul}$ ,  $VSM^{Quant}$ ,  $VSM^{ML}$  and  $VSM^{MAP}$  were compared with distributions obtained using a reference PSF. Second, 3D dose distributions were computed in a water phantom for the VSMs and for the reference PSF, below the flattening filter and without considering the collimation systems. Third, the comparison of the 3D dose distributions was performed in a water tank and a heterogeneous phantom, after collimation systems.

**Comparison in the phase space plane.** The different VSMs were used to generate about  $5.10^9$  particles which were transported in PENELOPE until they crossed the phase space plane located at 16 cm from the electron source. A reference PSF (240 Go) storing three times more photons than the PSF used to construct the VSMs was computed at the same location. The energy, polar and azimuthal distributions of the particles stored in the PSF files were generated for the  $VSM^{Reg}$ ,  $VSM^{Cumul}$ ,  $VSM^{Quant}$ ,  $VSM^{ML}$  and  $VSM^{MAP}$ , and were compared to those obtained from the reference PSF, for each sub-source.

Two methods were used for the comparison of distributions. The first method consisted in calculating  $D_s$ , the sum of the square difference between the bins of the reference PSF distribution and those of the VSM distribution as defined by equation 4.  $N_{bin}$  denotes the number of bins in the distributions,  $d_{i,PSF,s}$  and  $d_{i,VSM,s}$  the values of the  $i_{th}$  bins in the distributions generated with the reference PSF and the VSM for the sub-source  $s$ , respectively.

$$D_s = \sqrt{\sum_{i=0}^{N_{bin}} (d_{i,PSF,s} - d_{i,VSM,s})^2} \quad (4)$$

The second method calculates the Hellinger distance  $H_s$ , suitable for probability distribution comparison, between the bin values for the reference PSF and the VSM distributions, for each sub-source of the VSM. This quantity is defined by equation 5:

$$H_s = \frac{1}{2} * \sqrt{\sum_{i=0}^{N_{bin}} (\sqrt{d_{i,PSF,s}} - \sqrt{d_{i,VSM,s}})^2} \quad (5)$$

**Comparison of dose calculations in water below the flattening filter.** Dose distributions within a water phantom were computed separately for each sub-source, and for the VSM combining the three

sub-sources. The water phantom had dimensions of  $60 \times 60 \times 30 \text{ cm}^3$  for all calculations and was located at a source-to-surface distance (SSD) of 17 cm. The voxel size was  $0.1 \times 1 \times 1 \text{ cm}^3$  for the lateral dose profiles and  $1 \times 1 \times 0.3 \text{ cm}^3$  for the depth dose profiles (DDP). A number of  $2 \cdot 10^9$  particles were generated for each sub-source using the VSMs and the reference PSF. DDPs, as well as lateral dose profiles at 1.5, 5 and 10 cm water depth obtained with the reference PSF and the VSMs were compared using a global gamma index. A distance criterion of 0 mm was set since no distance error exists in the simulation results. The global dose criterion was set to 1.5% for the lateral profiles and to 1% for the DDPs, respectively. It was chosen so that it corresponds to the uncertainty observed in simulated dose profiles at the maximum dose point. These comparisons helped us to verify the accuracy of our model in a wide range of radial positions. The Russian roulette was here disabled.

**Comparison of dose calculations after the collimation systems.** The next validation step involved dose calculations both in the same water phantom than used previously, and located at a SSD of 100 cm, and in a layered heterogeneous phantom whose entrance face was located at a SSD of 90 cm. This latter phantom was composed of a first  $30 \times 30 \times 4 \text{ cm}^3$  slab of water equivalent material ( $1.04 \text{ g/cm}^3$ ), a second  $30 \times 30 \times 2 \text{ cm}^3$  slab of CIRS bone equivalent material ( $1.8 \text{ g/cm}^3$ ). A third layer was made of two  $30 \times 16 \times 10 \text{ cm}^3$  slabs of CIRS lung equivalent material ( $0.3 \text{ g/cm}^3$ ), separated by a 3 cm air gap. The fourth layer was made of another  $30 \times 30 \times 6 \text{ cm}^3$  slab of water equivalent material. Calculations for both phantoms were performed for a field size of  $10 \times 10 \text{ cm}^2$ , for the  $\text{VSM}^{\text{Cumul}}$ ,  $\text{VSM}^{\text{Quant}}$ ,  $\text{VSM}^{\text{ML}}$  and  $\text{VSM}^{\text{MAP}}$ . Voxel sizes were set to  $1 \times 1 \times 0.3 \text{ cm}^3$  and  $0.1 \times 2 \times 2 \text{ cm}^3$  for DDPs and lateral profile calculations, respectively. A number of  $5.2 \cdot 10^{10}$  particles were generated for each calculation using the VSMs. DDPs as well as lateral dose profiles at 1.5, 5, 10 and 20 cm obtained with the reference PSF and the VSMs were compared using a global gamma index with 2%-0 mm criteria. The use of the VSM was restricted for the small field sizes by setting the Russian roulette parameter R to 7 cm and the particle survival probability  $P_{\text{RR}}$  to 0.7.

### 2.3.3. Application of VSMs to EPID image prediction.

The generic representation of the EPID described by Badel (Badel, 2009) was used to model the EPID (figure 4). A 2 cm water slab was added to model backscatter materials. Electron and photon threshold energy cutoffs were set to 300 keV for electrons/positrons, and 10 keV for photons in PENELOPE while  $C_1$  and  $C_2$  are set to 0.2, except in the gadox layer where they were equal to 0.1.  $W_{\text{cc}}$  and  $W_{\text{cr}}$  were respectively set to 300 and 10 keV. DSMAX is set to 0.01 cm in copper layer and 0.0054 cm in gadox layer. These choices corresponded to the tenth of thickness layers.

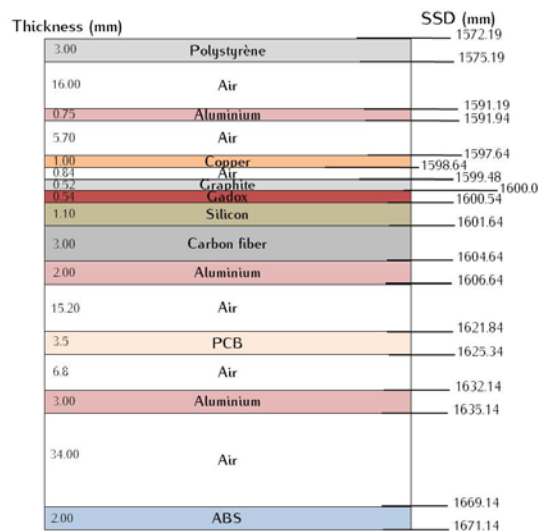


Figure 4. EPID modelling in PENELOPE.

EPID images were computed using the method described in Lazaro et al (Lazaro et al., 2013), for different configurations. The EPID was located at 160 cm from the target. Images were calculated with and without phantom in the beam for a  $10 \times 10 \text{ cm}^2$  field. Two kinds of phantoms were considered: a

40 cm thickness homogeneous water phantom located at a SSD of 100 cm, and the same layered heterogeneous phantom as described in the previous section, whose entrance face was located at a SSD of 90 cm.

Calculations of EPID images were performed for the  $VSM^{Cumul}$ ,  $VSM^{ML}$  and  $VSM^{MAP}$ . Voxel sizes were set to  $0.04 \times 0.04 \times 0.054 \text{ cm}^3$  to reproduce real EPID ones. A number of photons equivalent to  $1.10^9$  primary electrons were generated for each calculation using the VSMs. As before, the Russian roulette parameter R and  $P_{RR}$  were respectively set to 7 and 0.7. Lateral profiles and 2D-images obtained with the reference PSF and the VSMs were compared using a global gamma index with 1% / 0 mm criteria.

### 3. Results and discussion

#### 3.1. Physical interpretation of correlated histograms and analysis of correlations

1D and 2D-correlated histograms were first examined, and correlations between variables were studied on 2D histograms.

**3.1.1. 1D-histograms.** Sorting of particles in the PSF showed that 69%, 8% and 23% of the particles last interacted in the target (T), the primary collimator (PC), and the flattening filter (FF), respectively. These contributions correspond to the intensities of the three sub-sources in our VSM. Figure 5 shows the superposition of the radial, energy, polar and azimuthal angle distributions obtained for each sub-source separately, and for all particles. The contribution of each sub-source was weighted by its intensity. Distributions are represented with a regular binning, with bins of 0.2 mm, 5 keV and 0.5 deg.

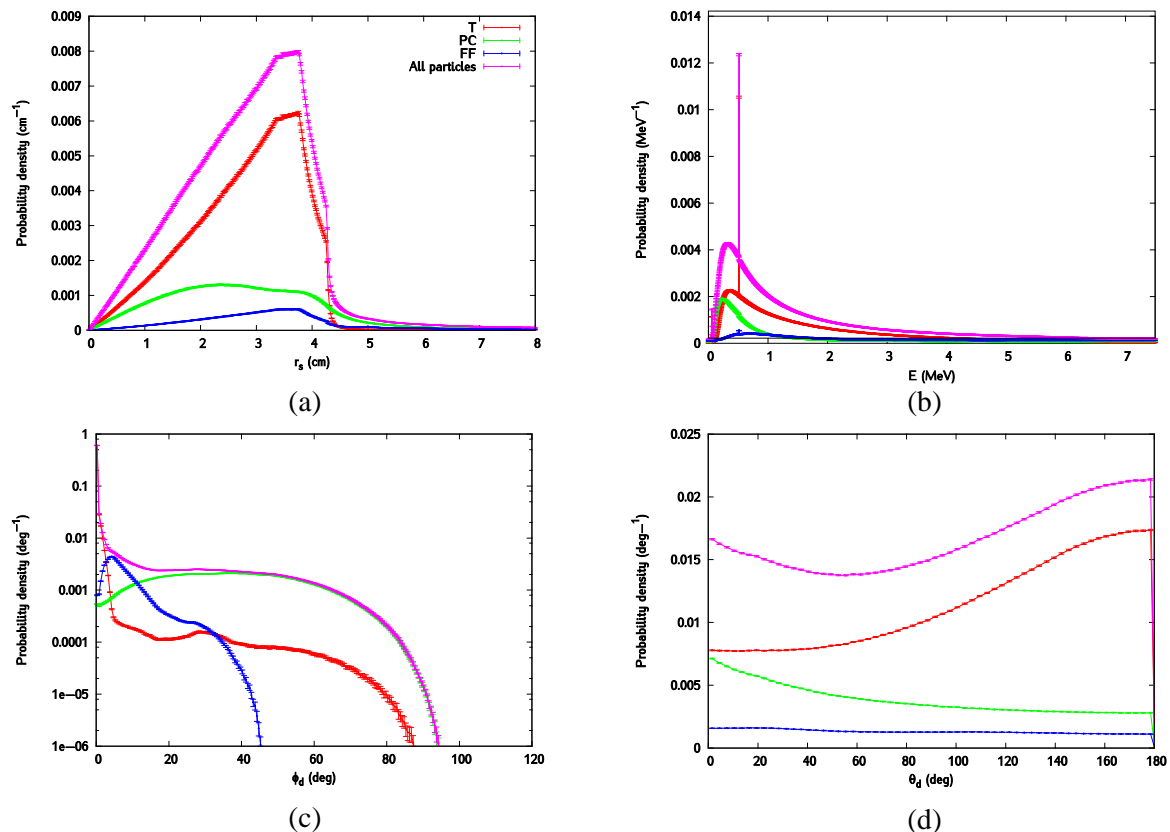


Figure 5. Distributions of (a) radial position, (b) energy, (c) polar angle, and (d) azimuthal angle for the T, the PC, the FF sub-sources, and for all sub-sources (Tot).

Radial position distributions (figure 5a) show an expected increase in particle number due to the constant width of the bins. Indeed, the rings delimited by the radial bins have an increasing area while

moving from the central axis to off-axis, which in turn increases the particle probability. Radial position distributions are clearly dependent on both the linac geometry and the sub-source. Figure 5b exhibits same trends for energy distributions than those already observed in previous studies (Schach von Wittenau et al., 1999). Polar angle distributions (figure 5c) show a strong dependence on the type of the sub-source. The polar angle distribution for the T sub-source is nearly a Dirac delta function. The width of the observed peak is due to particles interacting in the target before leaving it and which are emitted all along the target. To the contrary, the polar angle distributions for the PC and FF sub-sources are poorly filled at small angles. At wide angles, the FF distribution clearly dominates others, showing that multiple scatter is possible in the FF before particles leave it, and that it could result in polar angle values which may be high. Particles interacting in the PC must scatter at wide angles to be able to leave it. Moreover, multiple scatters inside the PC decrease the chance of the particle to go out. Azimuthal angle distributions are more difficult to interpret since the coordinate system is mobile.

3.1.2. *2D-histograms.* Figure 6 shows the example of the 2D-histograms representing the radial position distribution of particles according to their energy for the T (figure 6a), the PC (figure 6b), the FF (figure 6c) sub-sources, and the combination of all of them (figure 6d).

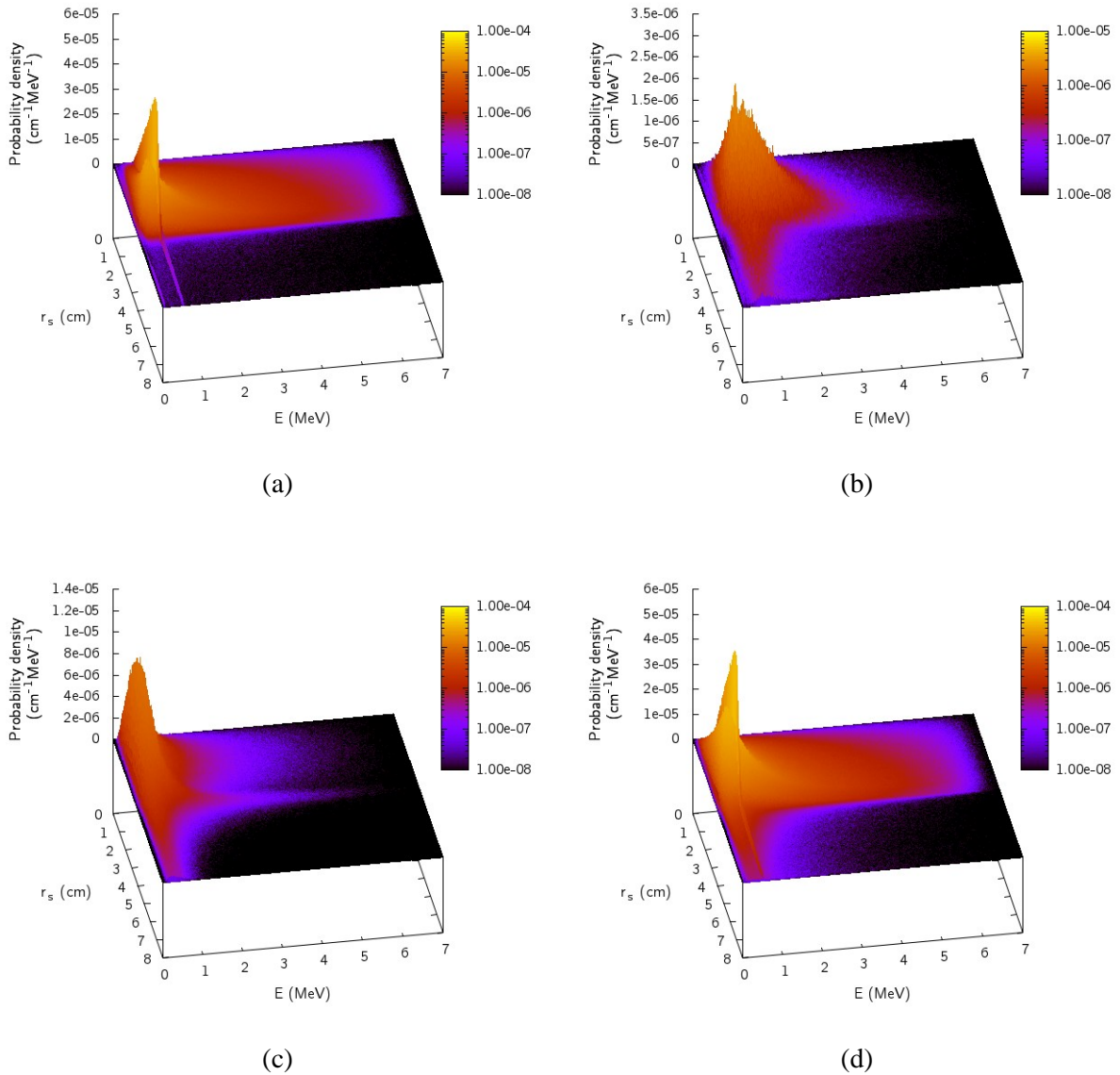


Figure 6. 2D-histograms showing the evolution of the energy distribution according to the radial position of particles for the (a) T, (b) PC, (c) FF sub-sources, and (d) for all particles.

For all sub-sources, particles of high-energy were always observed close to the PC edges, around a radial position of 3.5 cm. This is mainly due to the fact that the bin area and the number of particles increase both with the radial position. High-energy particles were most likely observed on-axis. Each sub-source distribution presented however specificities. The T sub-source particles were stopped by the PC and their radial positions were smaller than 3.5 cm. Only few annihilation photons were observed. Most of the photons interacting with the PC had a radial position less than 3.5 cm but a significant number of low energy particles were however observed beyond. The shape of the FF sub-source distribution was less impacted by the PC than the others.

Figure 6 shows the 2D-histograms corresponding to the product of marginal distributions that could have been observed if no correlation had existed between the radial position and energy distributions. It can be clearly seen that they are different from the joint distributions represented on figure 7. The same observations were made for the other variables compared two-by-two.

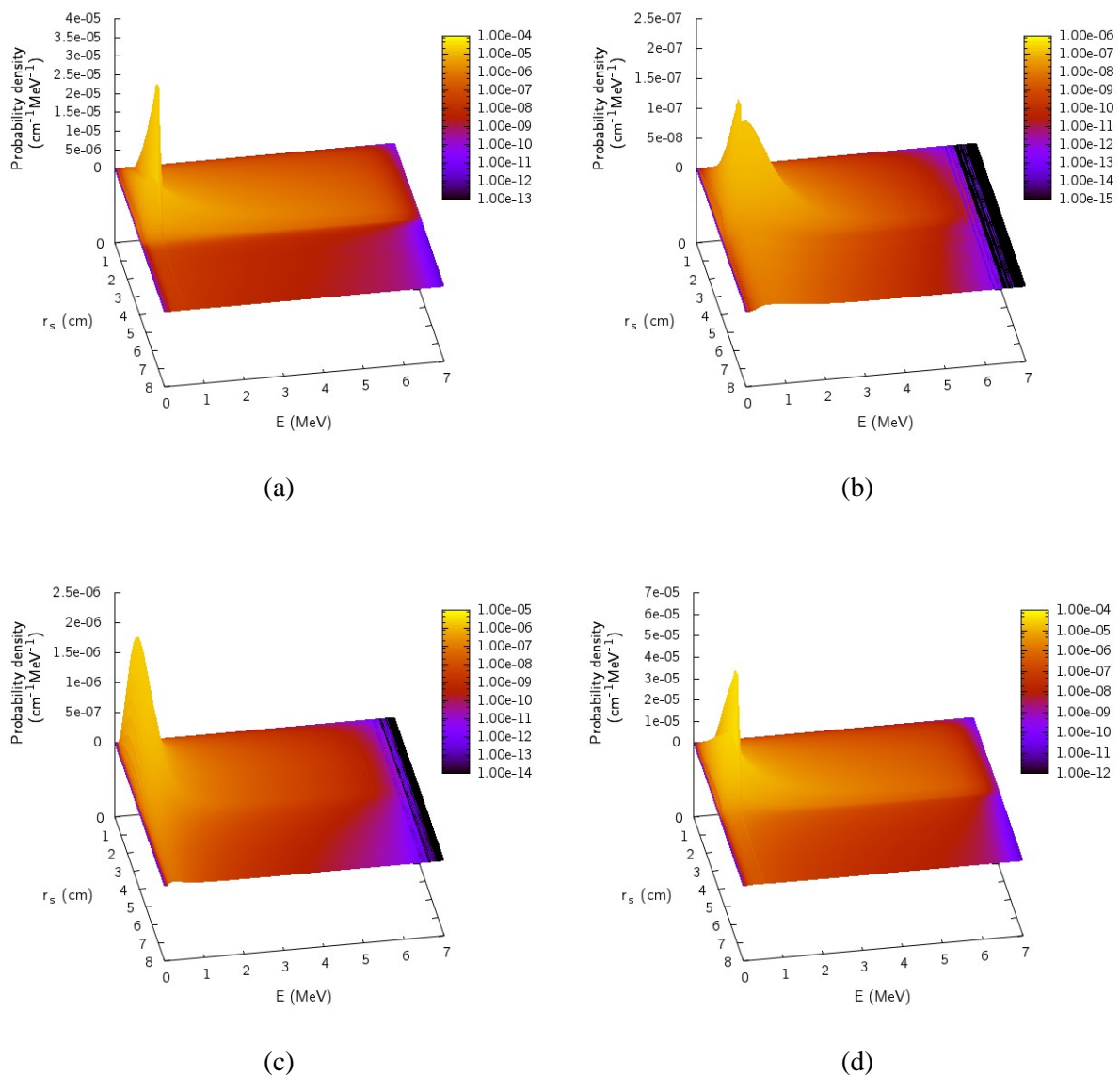


Figure 7. Product of marginal distributions showing the evolution of the energy distribution according to particle radial positions for the (a) T, (b) PC, (c) FF sub-sources, and (d) all particles, if no correlation had existed between the radial position and the energy.



The results of the Hoeffding test, summarized in table 2, show that the observed differences between the joint distributions and the product of marginal distributions are significant, and hence confirm that correlations exist between all of the four variables  $r_s$ ,  $E$ ,  $\varphi_d$ ,  $\theta_d$ , and for each sub-source of the VSM.

**Table 2.** Results of the Hoeffding independence test for the three sub-sources of the VSM. The threshold for the p-values was set to 0.01.

Sub-source	2D-histograms						
	$r_s, E$	$r_s, \varphi_d$	$r_s, \theta_d$	$E, \varphi_d$	$E, \theta_d$	$\varphi_d, \theta_d$	
T	H	4,78E-01	1,08E+00	8,19E-01	4,78E+00	9,92E-01	3,40E+00
	p	1,23E-07	6,90E-11	9,49E-10	< 1E-16	1,52E-10	5,55E-16
PC	H	1,28E+00	3,80E+00	3,29E+00	1,21E+01	5,87E+00	9,33E+00
	p	1,25E-11	2,22E-16	7,77E-16	< 1E-16	< 1E-16	< 1E-16
FF	H	1,23E+00	1,60E+00	4,92E+00	1,57E+01	1,48E-01	5,60E-01
	p	1,83E-11	1,36E-12	< 1E-16	< 1E-16	6,79E-04	3,07E-08

Results suggest that it is hence worth trying to keep all these correlations in the VSM if one wants to preserve most of the accuracy stored in the PSF.

### 3.2. Performances of the VSM

#### 3.2.1. Study of the binning: comparison in the phase space plane.

the comparison of the distributions obtained from the VSM<sup>MAP</sup> and from the reference PSF. Distributions were expressed in the linac coordinate system.

Figure 8 shows as an example

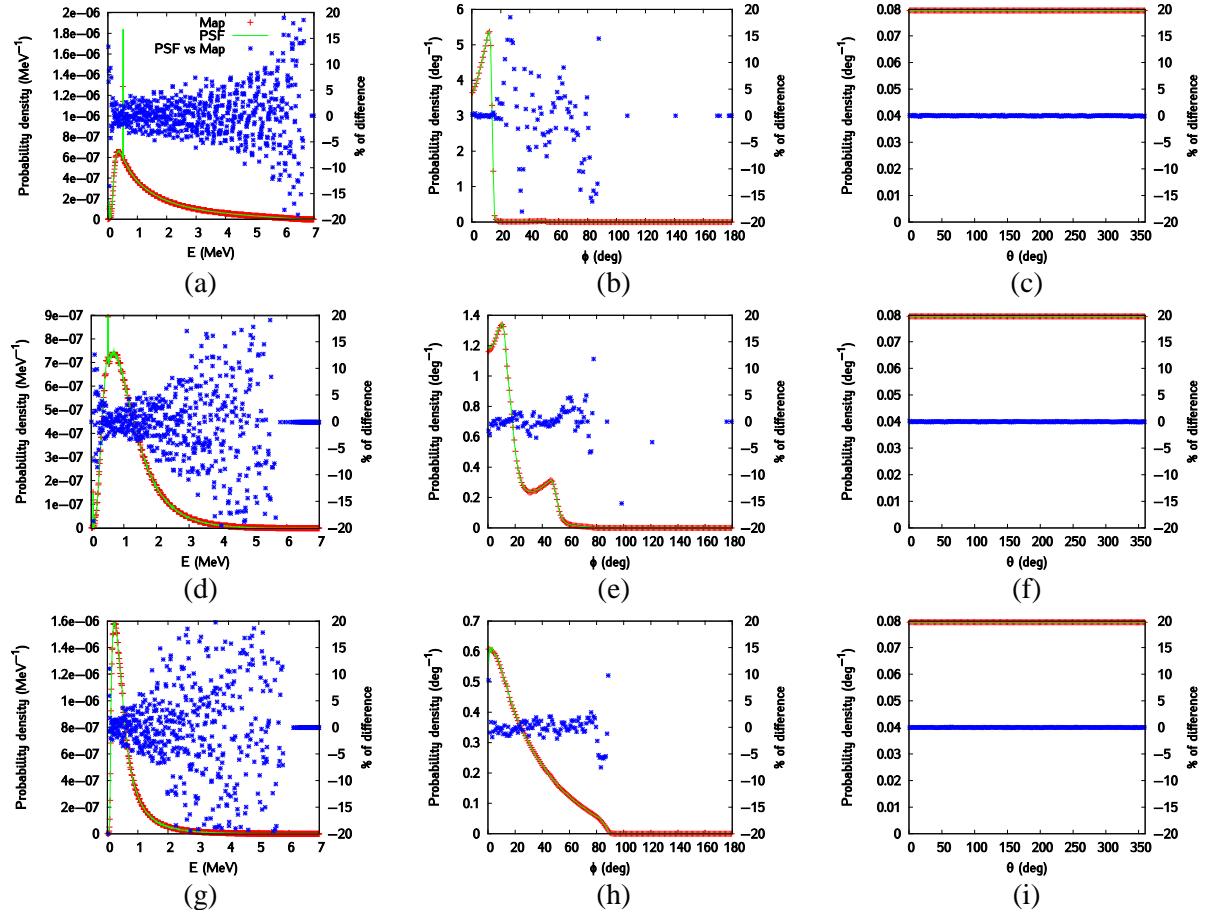
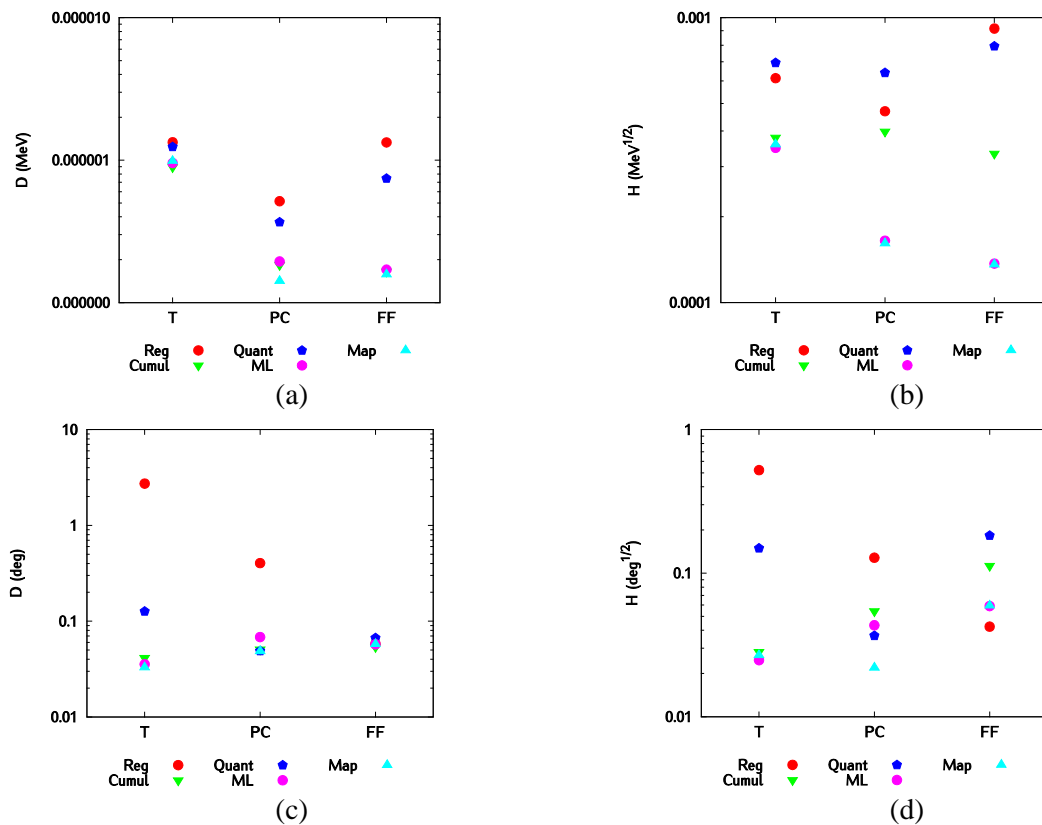


Figure 8. Comparison of the energy (left), the polar angle (middle) and the azimuthal angle (right) distributions obtained from the  $VSM^{MAP}$  (red points) and from the reference PSF (green line), for the T (first line), the PC (second line) and the FF (third line) sub-sources. Blue points represent the relative difference between both distributions.

A very good agreement can be observed for all the distributions. We can see the impact of using an irregular binning on the energy distribution. Indeed, as the energy is sampled uniformly within a bin, it could be overestimated or underestimated when sampled at the inner edge or at the outer of the bin, respectively. The larger the bin, the greater the difference between the energy value at the inner and at the outer edges of the bin. Since bins become less and less filled when energy increases, hence larger and larger for the *MAP* binning, discrepancies between energy distributions from the VSM and from the reference PSF increase consequently. In general, larger discrepancies were observed for bins of small amplitude. The shape of the polar angle distribution depends on the PSF radial value. If the PSF is large, the second fall of the distribution would be shifted towards the highest angle values. The azimuthal angle distributions are uniform, and the axial symmetry of the beam is thus satisfied for all sub-sources.

Figure 9 shows the comparison of the distributions for each binning scheme and for each sub-source separately, using the D and H distance criteria. The lower the calculated value of D and H, the lower the discrepancy between the calculated distributions from the VSM and from the reference PSF.





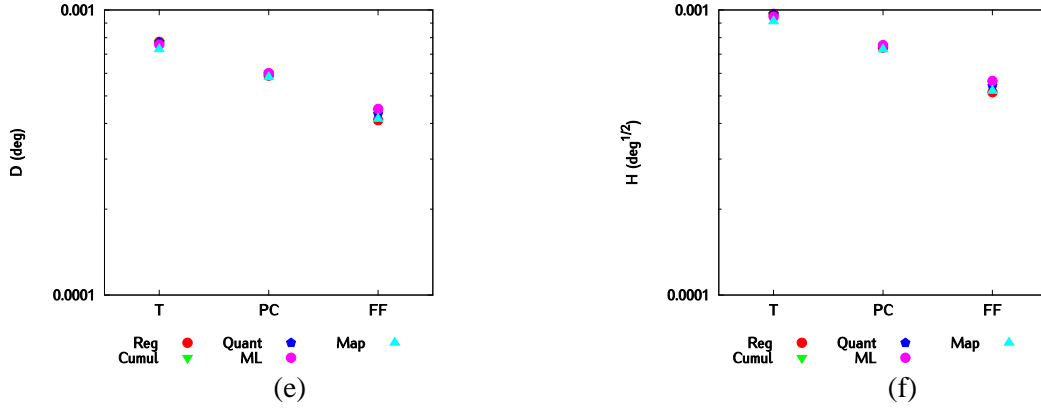


Figure 9. Comparison between the five studied binning schemes using the H and D distance criteria for the three sub-sources. (a) Energy distributions – Distance D, (b) Energy distributions – Distance H, (c) Polar angle distributions – Distance D, (d) Polar angle distributions – Distance H, (e) Azimuthal angle distributions – Distance D, (f) Azimuthal angle distributions – Distance H.

The study of  $D_S$  and  $H_S$  values show that energy and polar angle distributions are sensitive to the binning scheme. The *Reg* and *Quant* binning schemes lead to the greater values of  $D_S$  and  $H_S$ , and hence to the bigger discrepancies between the distributions. The other binning schemes, *Cumul* and those based on the estimation of maximization of the likelihood (*ML* and *MAP*), give similar values of  $D_S$  and  $H_S$ , smaller than those obtained with *Reg* and *Quant*. This is particularly true for energy distributions shown in figure 9b. The *Cumul* binning give contrasted results: they are as good as those obtained with *ML* and *MAP* for the target energy and the polar angle distributions, but discrepancies are more pronounced for the PC and the FF distributions. The azimuthal angle distributions figure 9e and figure 9d) were well reproduced by any binning scheme. Figure 10 illustrates the difficulty of the *Reg* binning to reproduce the energy and the polar angle distributions for the target sub-source. Indeed, the details present at low energies could not be correctly reproduced (figure 10a). Moreover, the use of the *Reg* binning changes greatly the shape of the polar angle distribution for the T sub-source, for small polar angle values.

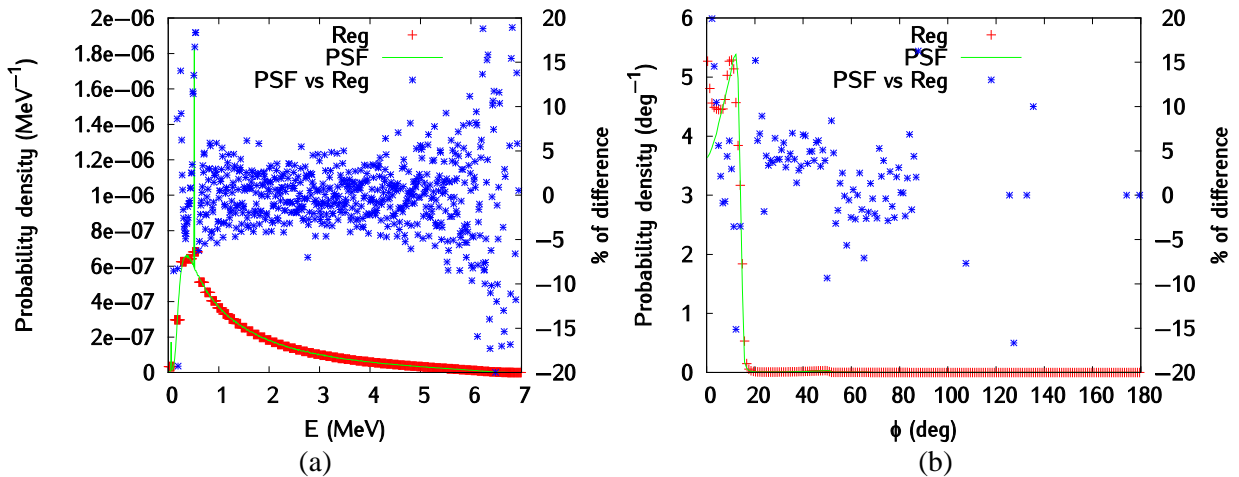


Figure 10. Comparison of (a) energy and (b) polar angle distributions obtained from the T sub-source of  $VSM^{Reg}$ , and from the reference PSF.

3.2.2. *Study of the binning: comparison of dose calculations below the FF.* All dose distributions were calculated with a statistical uncertainty ( $3\sigma$ ) lower than the gamma dose criterion. Figure 11 shows the lateral dose profiles calculated in the water phantom at 1.5, 5, 10 and 20 cm depth with the

VSM<sup>Reg</sup> for the T (figure 11a), the PC (figure 11b), and the FF (figure 11c) sub-sources. This sub-source by sub-source comparison shows very big discrepancies: dose profiles obtained for the VSM<sup>Reg</sup> are represented on figure 11d, and the comparison with the reference PSF confirms that the VSM constructed with the *Reg* binning is not able to supply accurate dose calculations.

Figure 12 shows the same dose distributions calculated with the VSM<sup>Quant</sup>. This VSM gives overall good results, except for the PC sub-source, where some discrepancies are noted on axis. However, dose calculations performed with the VSM<sup>Quant</sup> were insensitive to these discrepancies, due to the fact that the PC sub-source represents only 8% of the total number of photons in this VSM. These discrepancies do not affect DDPs, either.

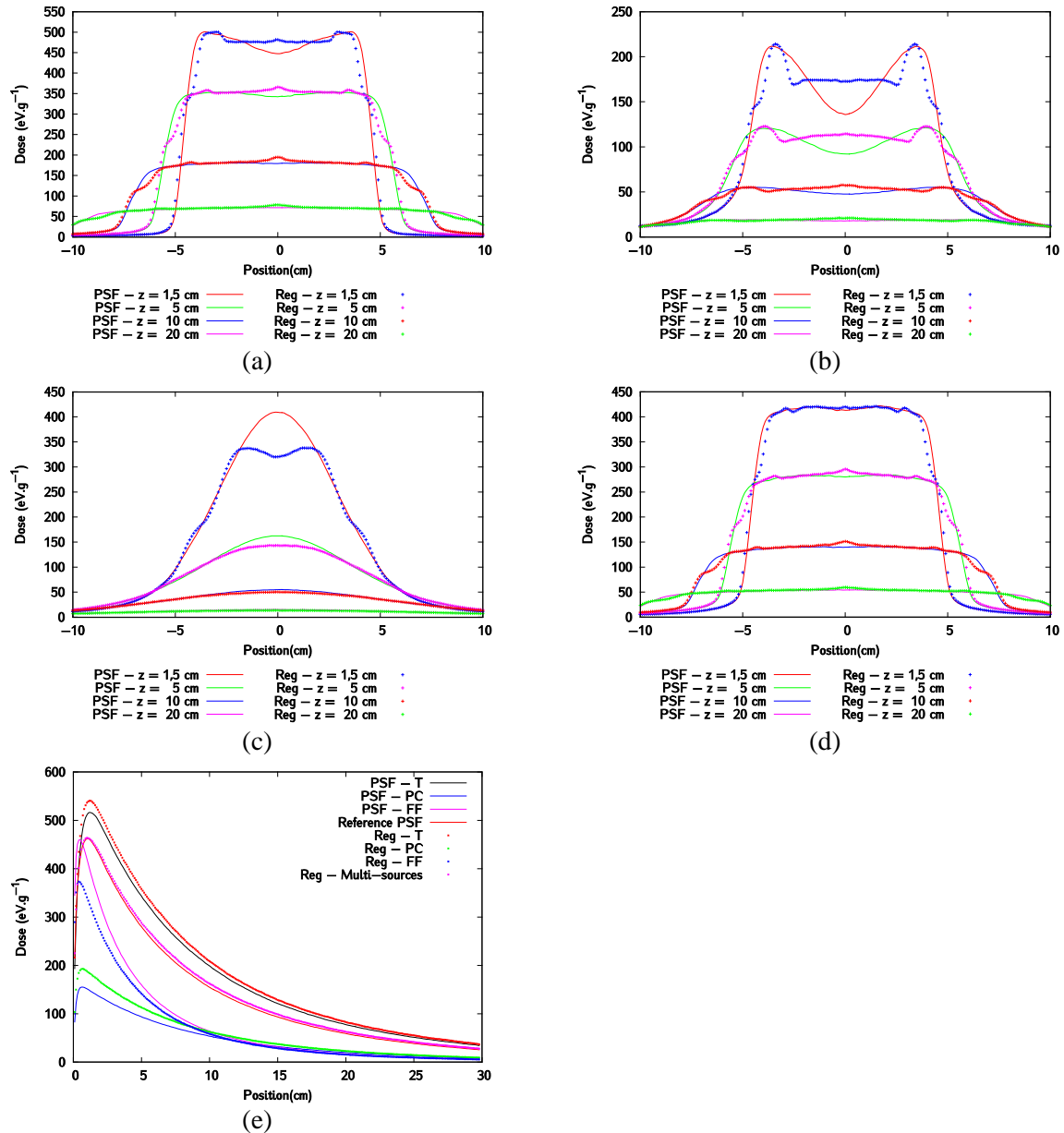


Figure 11. Comparison of lateral dose profiles calculated at 1.5, 5, and 10 cm depth with the VSM<sup>Reg</sup> (points) and with the reference PSF (solid line), for (a) the T, (b) the PC, (c) the FF sub-sources, (d) the VSM<sup>Reg</sup>, (e) Depth dose profiles.

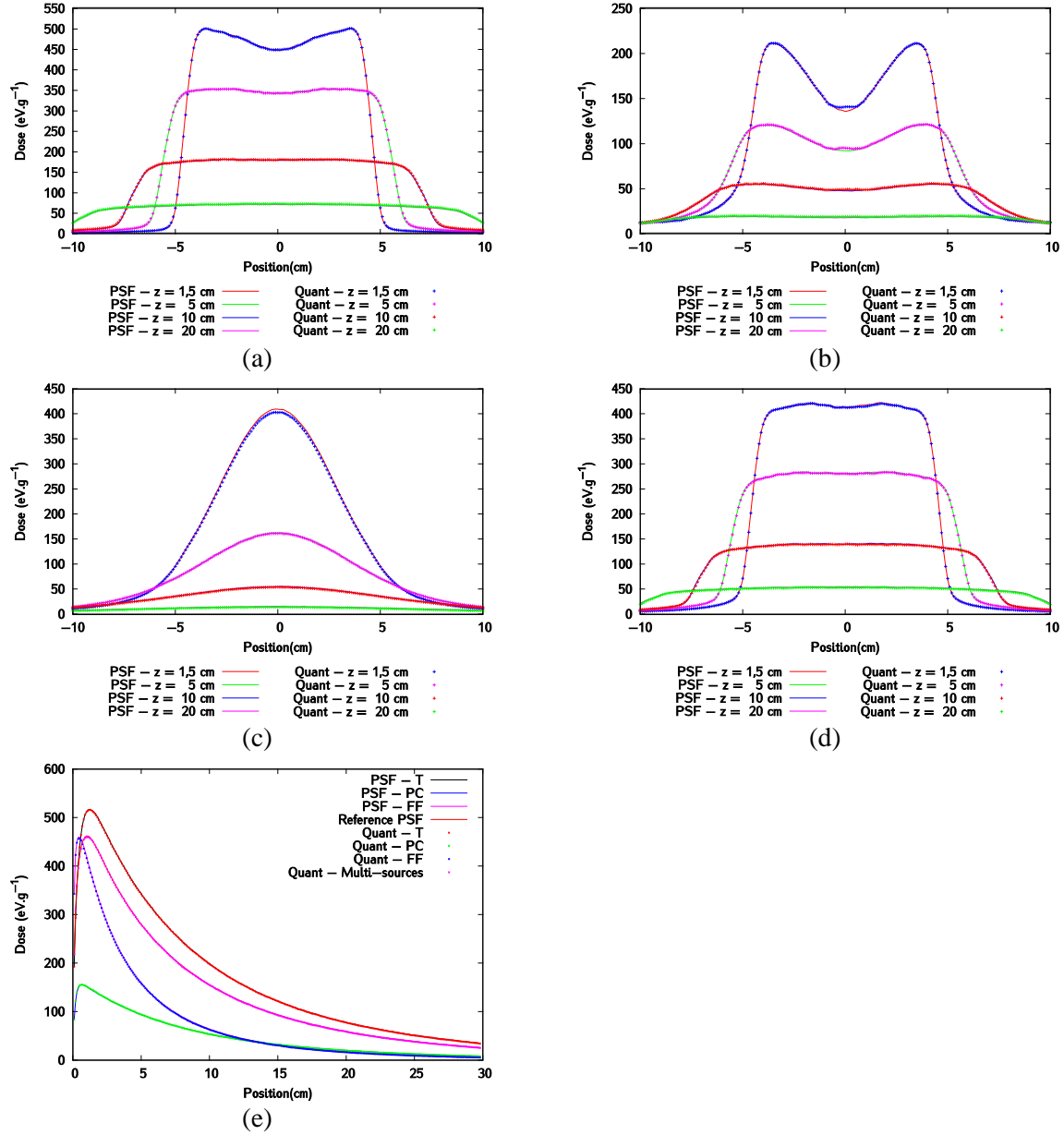


Figure 12. Comparison of lateral dose profiles calculated at 1.5, 5, and 10 cm depth with the  $\text{VSM}^{\text{Quant}}$  (points) and with the reference PSF (solid line), for: (a) the T, (b) the PC, (c) the FF sub-sources, and (d) the  $\text{VSM}^{\text{Quant}}$ . (e) Depth dose profiles.

The *Cumul*, *ML* and *MAP* binning schemes exhibit the same behaviour when used to construct the VSM. Figure 13 shows the lateral dose and depth dose profiles obtained with the *MAP* binning. A quantitative analysis using the gamma-index test is performed. All points of the lateral profiles successfully passed the gamma-index analysis using a 1.5% dose criterion for the  $\text{VSM}^{\text{Cumul}}$ ,  $\text{VSM}^{\text{ML}}$ , and  $\text{VSM}^{\text{MAP}}$ , and for all distributions. For DDPs, a 100% passing rate was obtained for all distributions. For the *Quant* binning, all distributions passed the gamma-index analysis with at least a 97% passing rate. For DDPs, points which failed were rather located in the build-up region or near this region.

Mono-source VSMs were also built for comparison with the multi-source VSMs studied until now. They are noted  $\text{VSM}^{\text{Quant,mono}}$ ,  $\text{VSM}^{\text{Cumul,mono}}$ ,  $\text{VSM}^{\text{ML,mono}}$ ,  $\text{VSM}^{\text{MAP,mono}}$ , when they were built using the *Quant*, *Cumul*, *ML* and *MAP* binning schemes. They were simply obtained by constructing a single 4D-histogram storing all the photons from the original PSF, without sorting them. The *Quant* binning

leads again to poorer results than others binnings. For the other binning schemes, the multi-source and mono-source VSMs gave similar results. Results obtained with the  $VSM^{MAP,mono}$  and the  $VSM^{MAP}$  are presented on figure 13d and figure 13e, respectively. They show that a very good agreement was observed for both  $VSM^{MAP}$  and  $VSM^{MAP,mono}$ . This last result exhibits that the use of an adaptive and irregular binning brings the possibility to build a simple and accurate source model for dose distribution calculations. This simpler mono-source model allows to get rid of the creation of sub-sources.

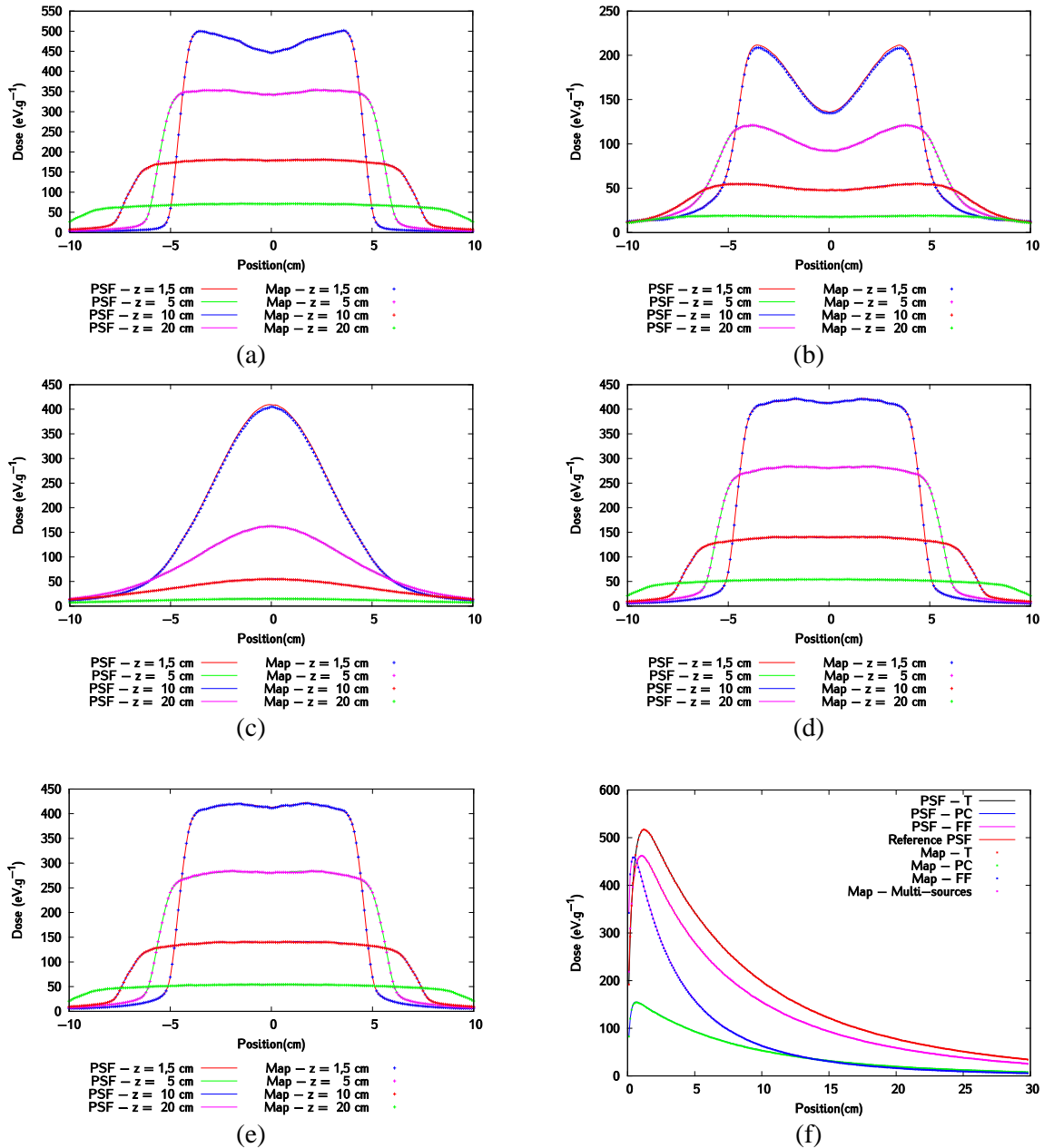


Figure 13. Comparison of lateral dose profiles calculated at 1.5, 5, and 10 cm depth with the  $VSM^{MAP}$  (points) and with the reference PSF (solid line), for: (a) the T, (b) the PC, (c) the FF sub-sources, (d) the mono-source  $VSM^{MAP,mono}$ , and (e) the multi-source  $VSM^{MAP}$ . (e) Depth dose profiles.

3.2.3. *Study of the binning: comparison of dose calculations after the collimation systems.* compare lateral dose profiles obtained for the  $VSM^{Quant}$ ,  $VSM^{Cumul}$ ,  $VSM^{ML}$ , and  $VSM^{MAP}$  with those from the reference PSF. shows the same comparison for the  $VSM^{Quant,mono}$ ,  $VSM^{Cumul,mono}$ ,  $VSM^{ML,mono}$ , and

VSM<sup>MAP,mono</sup> and compare DDPs obtained in the heterogeneous phantom with the multi-source and the mono-source VSMs, constructed using the *Quant*, *Cumul*, *ML* and *MAP* binning schemes.

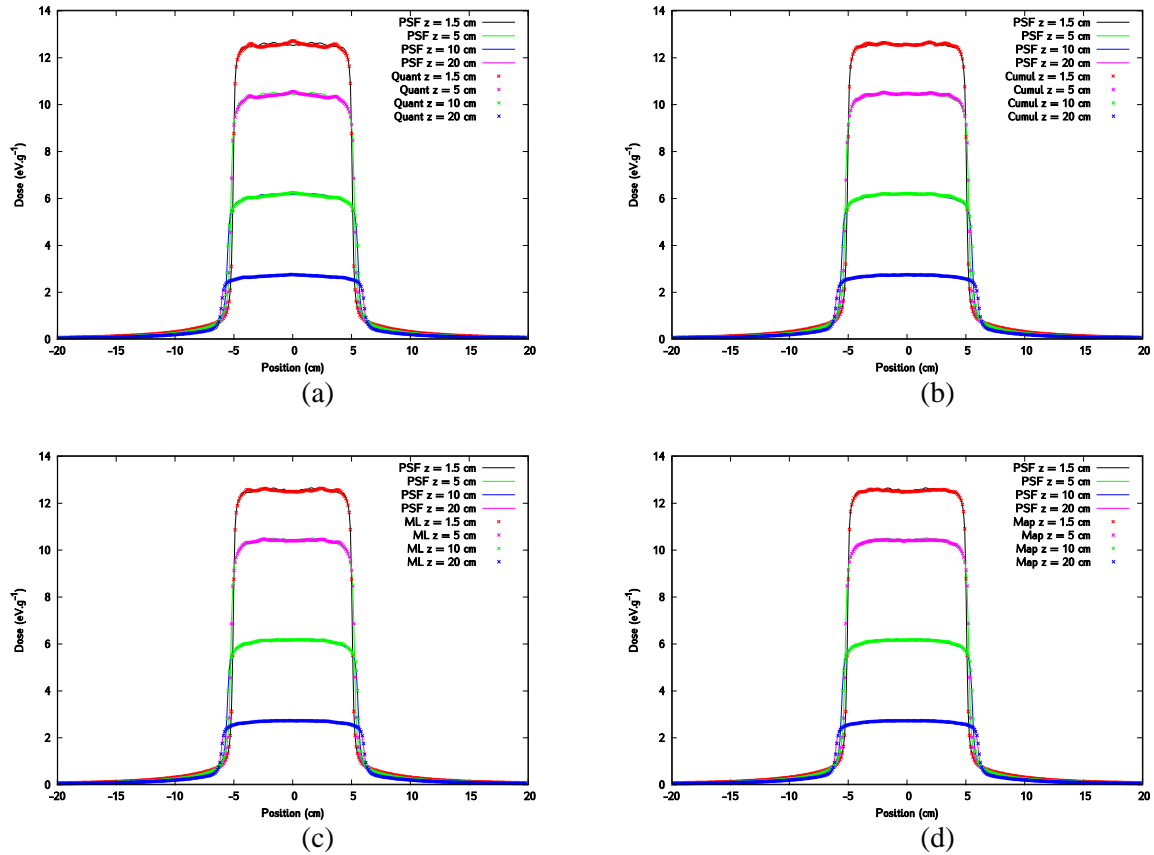
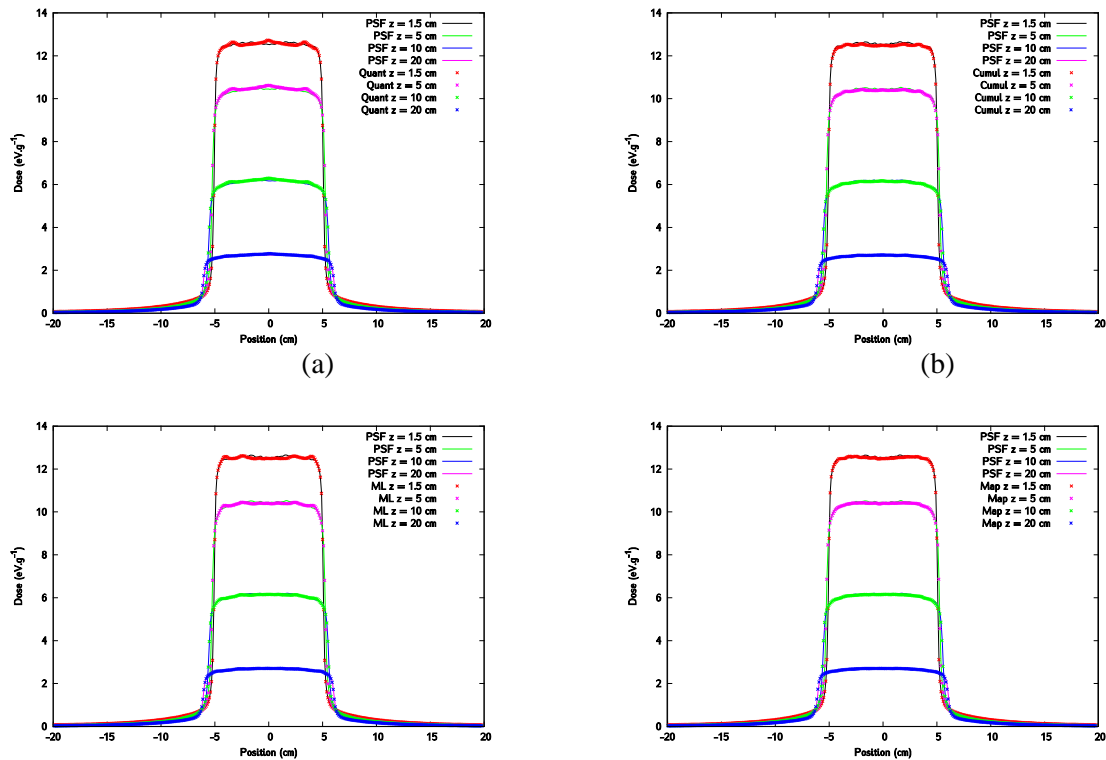
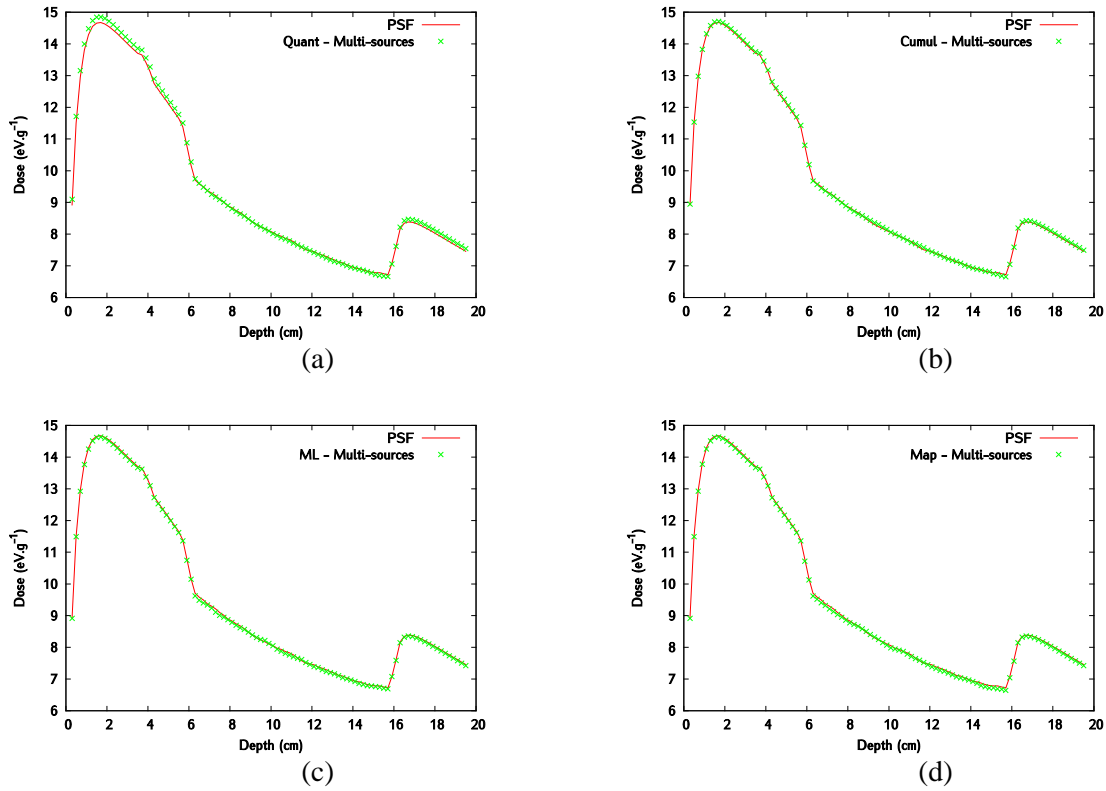


Figure 14. Comparison of lateral dose profiles calculated at 1.5, 5, 10 and 20 cm depth for (a) the VSM<sup>Quant</sup>, (b) the VSM<sup>Cumul</sup>, (c) the VSM<sup>ML</sup>, and (d) the VSM<sup>MAP</sup>.



(c) (d)  
 Figure 15. Comparison of lateral dose profiles calculated at 1.5, 5, 10 and 20 cm depth for (a) the  $VSM^{Quant,mono}$ , (b) the  $VSM^{Cumul,mono}$ , (c) the  $VSM^{ML,mono}$ , and (d) the  $VSM^{MAP,mono}$ .



(a) (b)  
 Figure 16. Comparison of depth dose profiles calculated in the heterogeneous phantom for (a) the  $VSM^{Quant}$ , (b) the  $VSM^{Cumul}$ , (c) the  $VSM^{ML}$ , and (d) the  $VSM^{MAP}$ .

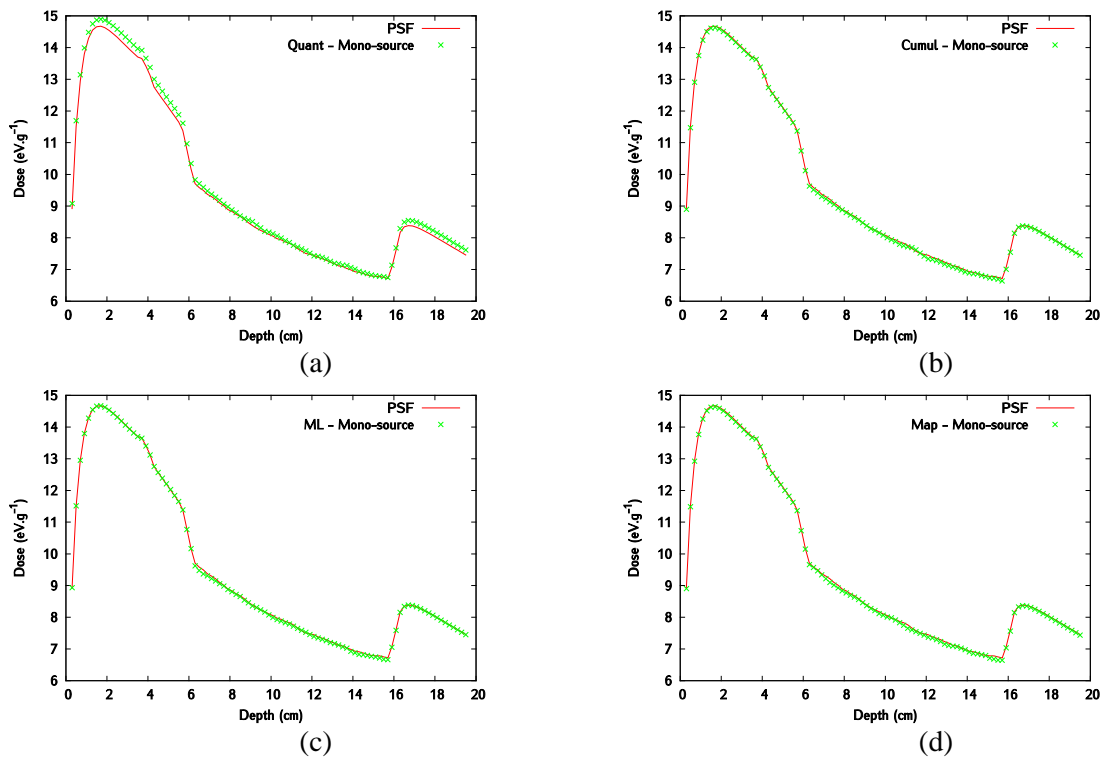


Figure 17. Comparison of depth dose profiles calculated in the heterogeneous phantom for (a)  $VSM^{Quant,mono}$ , (b) the  $VSM^{Cumul,mono}$ , (c) the  $VSM^{ML,mono}$ , and (d) the  $VSM^{MAP,mono}$ .

a and a show on the lateral dose profiles that the *Quant* binning leads to a peaked dose value on the axis, for both the  $VSM^{Quant}$  and the  $VSM^{Quant,mono}$ . This effect can also be clearly seen in the build-up region of the DDPs obtained for the heterogeneous phantom (figure 16a and figure 17a). All the other binning schemes, i.e. *Cumul*, *ML* and *MAP*, give very good results, both the multi-source VSMs and the mono-source VSMs. The comparison of the success rate  $\Gamma_{res}$  and of the mean value  $\Gamma_{moy}$  obtained for a 1 % / 0 mm criteria for both types of VSMs constructed using the different binning schemes is presented on the figure 18.

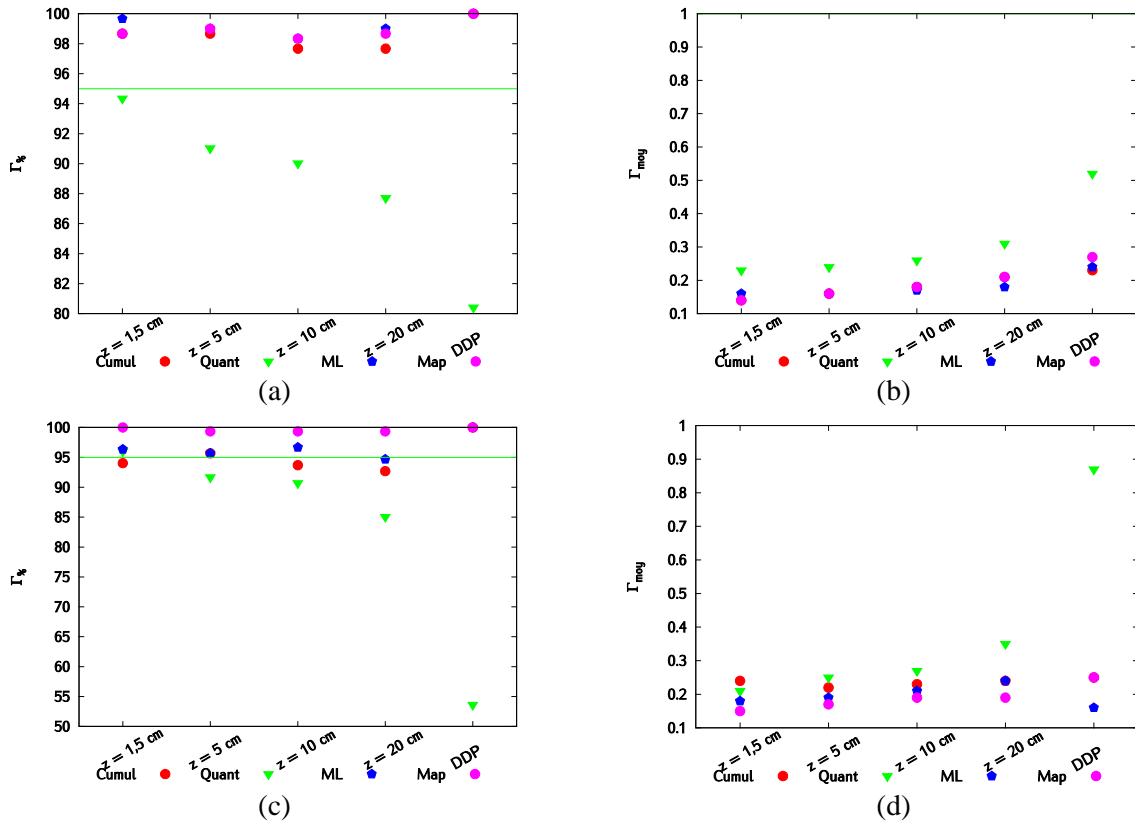


Figure 18. Comparison of (a)  $\Gamma_{res}$  and (b)  $\Gamma_{moy}$  calculated on lateral dose profiles and depth dose profiles for the multi-source VSMs ( $VSM^{Quant}$ ,  $VSM^{Cumul}$ ,  $VSM^{ML}$  and  $VSM^{MAP}$ ) and (c)  $\Gamma_{res}$  and (d)  $\Gamma_{moy}$  for the mono-source VSMs ( $VSM^{Quant,mono}$ ,  $VSM^{Cumul,mono}$ ,  $VSM^{ML,mono}$  and  $VSM^{MAP,mono}$ ).

The  $\Gamma_{res}$  and  $\Gamma_{moy}$  values obtained for the *Quant* binning are worse than for the other binning schemes. These results confirm that the *Quant* binning is not the appropriate binning scheme to construct a mono- or multi-source VSM. For the multi-source VSM, either the *Cumul*, *ML* or *MAP* binning schemes can be used to build the VSM, since they give all three very good and similar results. For the mono-source VSM, the *MAP* binning seems however to provide more robust dosimetric results than *Cumul* and *ML*.

### 3.2.4. Study of the binning: comparison of portal images.

Image prediction without phantom in the beam is challenging: in this case, the fluence calculated at the top of the EPID is very sensitive to variations in the radial and angular distributions of the VSM. This is accentuated by the small size of the EPID pixels, which is equivalent to 0.025 cm at the isocentre,

but also by the absence of any attenuating and diffusing media of a phantom/patient. Because the *Quant* binning scheme yields to poor results in the previous section, portal images obtained with it are not presented here. Crossplane profiles from portal images obtained using the reference PSF as well as *Cumul*, *ML* or *MAP* binning schemes with both mono- and multi-source VSMs are presented on figure 19. Agreements between the reference PSF and the mono-source models  $VSM^{Cumul,mono}$ ,  $VSM^{ML,mono}$  or the multi-source models  $VSM^{Cumul}$ ,  $VSM^{ML}$  are visually not satisfactory, as waves deform the profiles. These waves are due to bias introduced when we supposed a uniform bin density. In fact, a  $10 \times 10$  cm<sup>2</sup> field at the isocentre corresponds to about  $1 \times 1$  cm<sup>2</sup> at the PSF plane. We recall that the reference PSF was calculated for a maximum radial distance of 20 cm. Hence, only a few central bins of the VSM are involved for the image computation. This “zooming” effect tends to highlight discontinuities between two adjacent bins, which takes the shape of waves on the profiles. The 1% / 0 mm global gamma index analysis displayed on the figure 20 confirms that *Cumul* and *ML* binning schemes are less adapted to portal image prediction than *MAP* binning.  $\Gamma_{res}$  is indeed less than 95% for  $VSM^{Cumul,mono}$ ,  $VSM^{ML,mono}$  and  $VSM^{Cumul}$ . It reaches 97% in the case of  $VSM^{ML}$  which could be acceptable. The *MAP* binning scheme is more powerful since  $\Gamma_{res}$  is then greater than 99% whether the mono- or multi-source VSMs is used. Gamma index 2D maps corresponding to the comparison between the referenced PSF and  $VSM^{MAP,mono}$  and  $VSM^{MAP}$  are displayed on Figure 21. We can see that only few penumbra points fail the test. The slight discrepancy between mono- and multi-source models for each binning scheme could be explained by the prevalence of photons coming from the target sub-source in the central region of the beam.

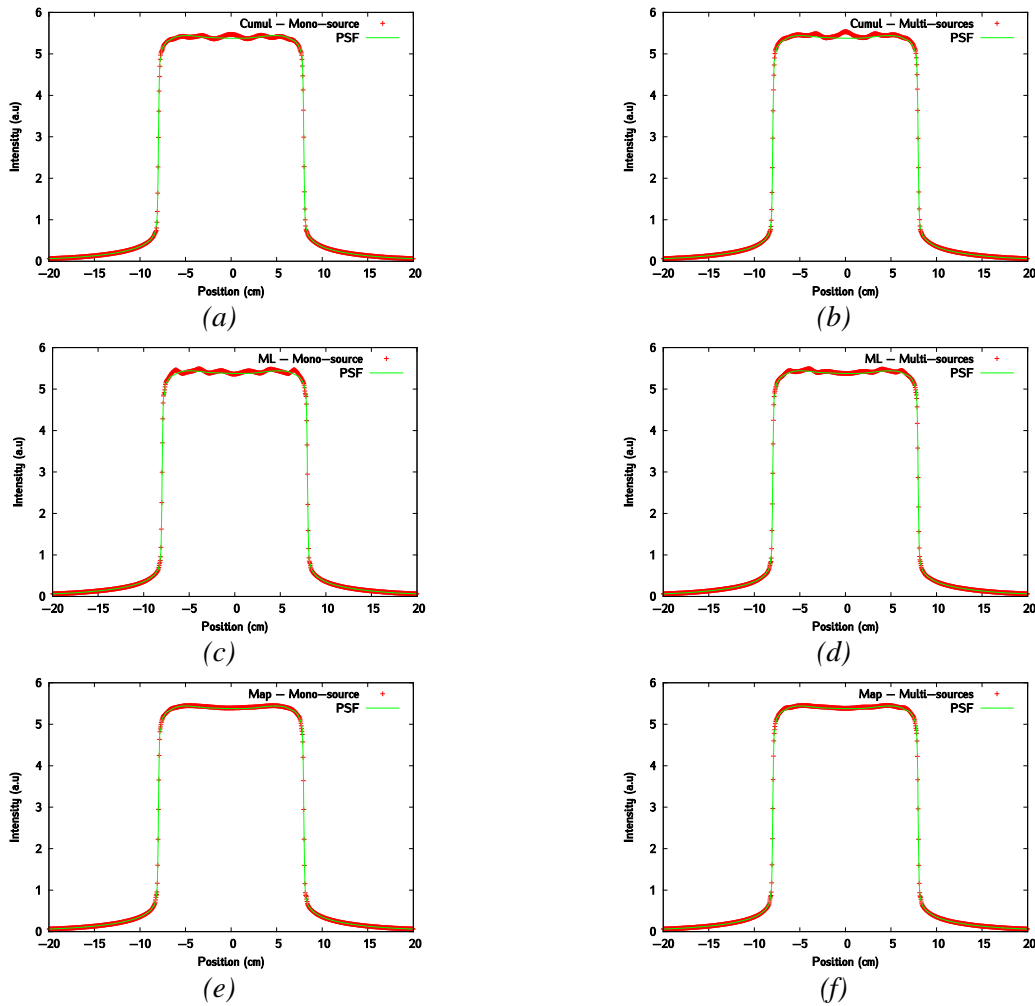


Figure 19. No phantom - Crossplane profiles of a  $10 \times 10$  cm<sup>2</sup> field obtained with the reference PSF, the mono-source models (a)  $VSM^{Cumul,mono}$ , (c)  $VSM^{ML,mono}$  and (e)  $VSM^{MAP,mono}$ , and the multi-sources models (b)  $VSM^{Cumul}$ , (d)  $VSM^{ML}$  and (f)  $VSM^{MAP}$ .



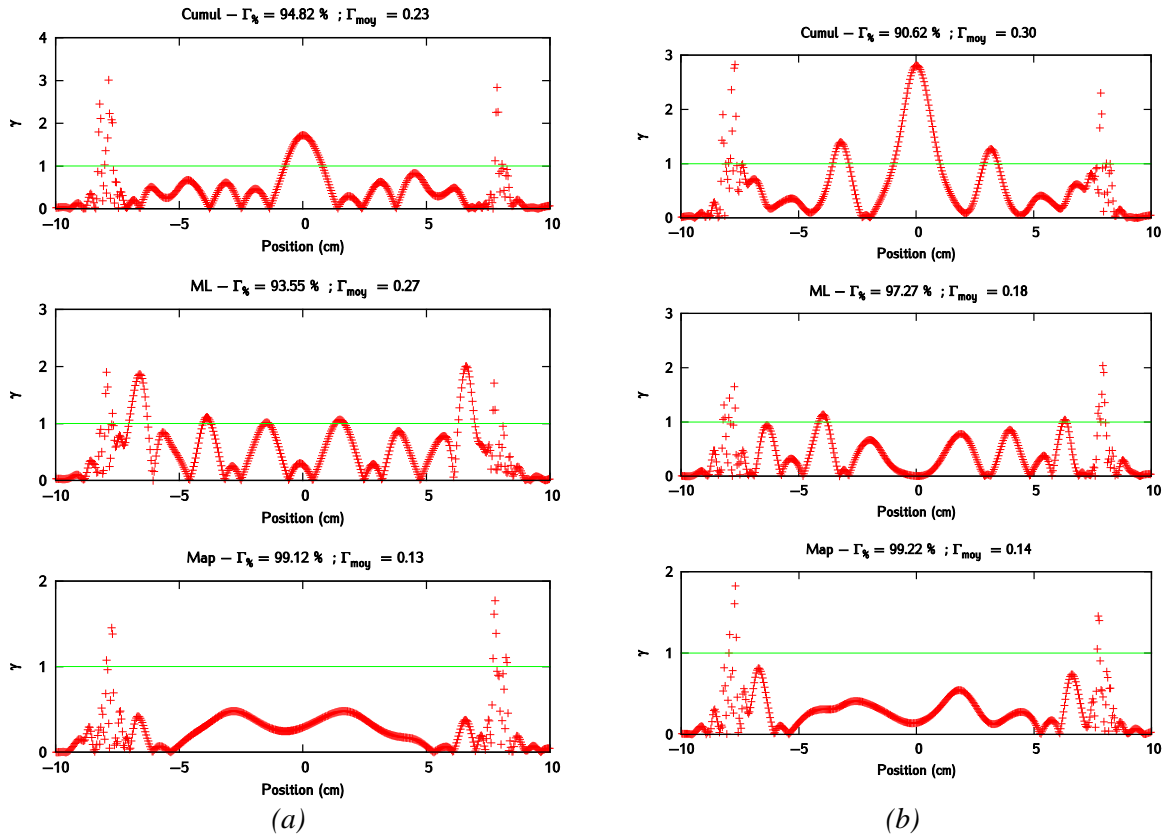


Figure 20. No phantom - Comparison of crossplane profiles of a  $10 \times 10 \text{ cm}^2$  field obtained with the VSM (a) the mono-source models and (b) the multi-source models and reference PSF using a global 1% / 0 mm gamma index.

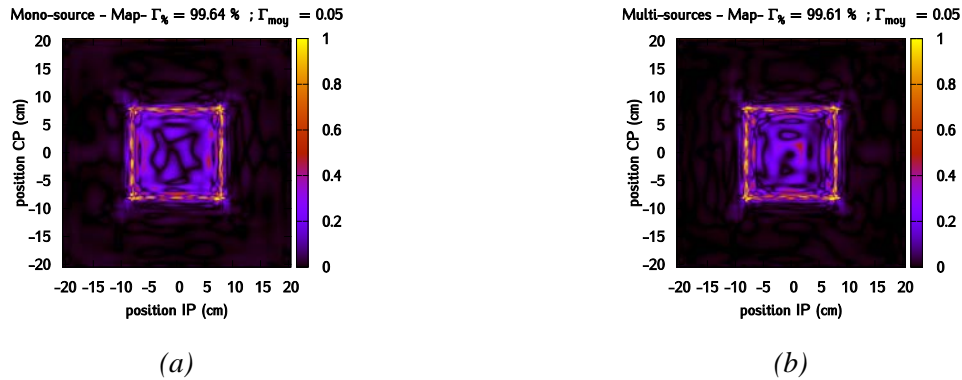


Figure 21. No phantom - Comparison of portal images of a  $10 \times 10 \text{ cm}^2$  field obtained with the VSM (a)  $\text{VSM}^{\text{MAP,mono}}$  and (b)  $\text{VSM}^{\text{MAP}}$  and reference PSF using a global 1% / 0 mm gamma index.

When the irradiation beam crosses a 40 cm thick water phantom, details of radial and angular distributions are blurred because of diffusion and attenuation effects. This configuration is nevertheless interesting since beam spectral properties are greatly changed compared to the previous section. Here are only presented results obtained with the *MAP* binning scheme since it led previously to the best results. The analysis of figures 22 and 23 leads us to the conclusion that the image prediction is excellent on the plateau and in the penumbra regions, except at the end of this latter. This explains  $\Gamma_{res}$  values around 94% with both VSMs. These discrepancies might be interpreted by considering the noisy dose image computation after crossing 40 cm water. Indeed, the denoising

algorithm that was used might over-smooth the regions where the dose gradient exhibits high curvatures.

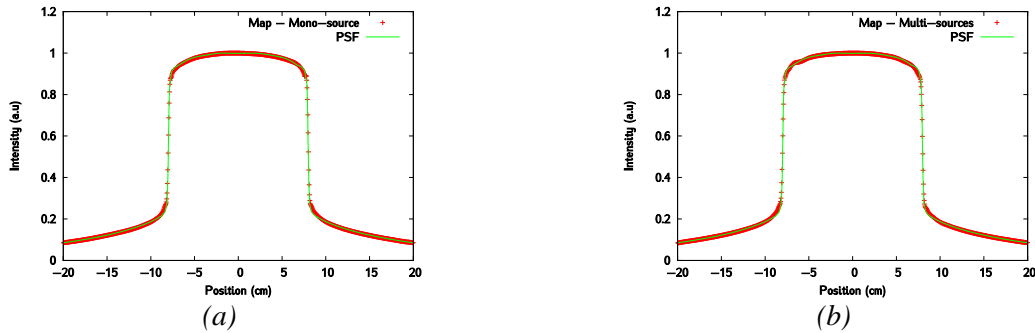


Figure 22. Water phantom - Crossplane profiles of a  $10 \times 10$  cm<sup>2</sup> field obtained with the reference PSF, the mono source model (a)  $VSM^{MAP,mono}$  and the multi-source model (b)  $VSM^{MAP}$

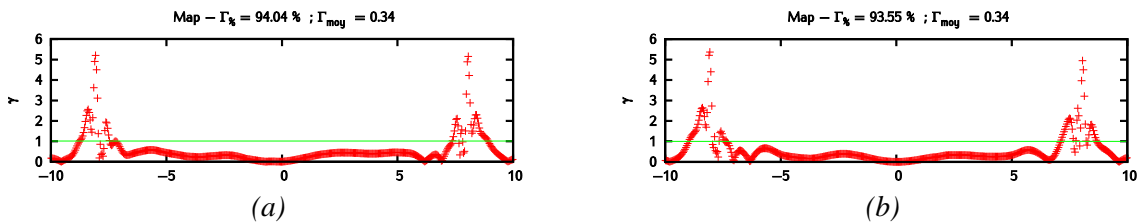


Figure 23. Water phantom - Comparison of crossplane profiles of a  $10 \times 10$  cm<sup>2</sup> field obtained with the VSM (a) the mono-source model  $VSM^{MAP,mono}$  and (b) the multi-source model  $VSM^{MAP}$  and reference PSF using a global 1% / 0 mm gamma index.

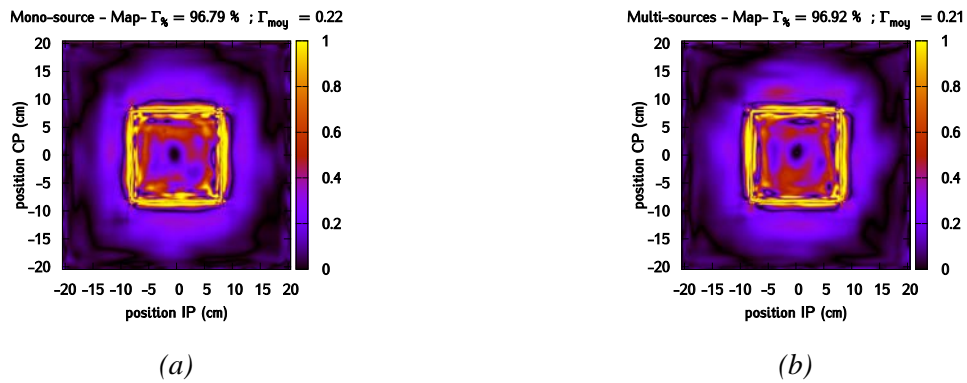


Figure 24. Water phantom - Comparison of portal images of a  $10 \times 10$  cm<sup>2</sup> field obtained with the VSM (a)  $VSM^{MAP,mono}$  and (b)  $VSM^{MAP}$  and reference PSF using a global 1% / 0 mm gamma index.

The very last validation step is the prediction of a portal image after the heterogeneous phantom. Here are tested the model ability of dose gradient restitution after material changes. Here again is the spectral distribution of the beam modified by the phantom. Figures 25 to 27 confirm good previous results obtained with the *MAP* binning scheme in both mono-source and multi-source models.  $\Gamma_{res}$  almost reaches 100% for the profiles while it is greater than 98% in the 2D maps.  $\Gamma_{res}$  is around 0.12 which is low.

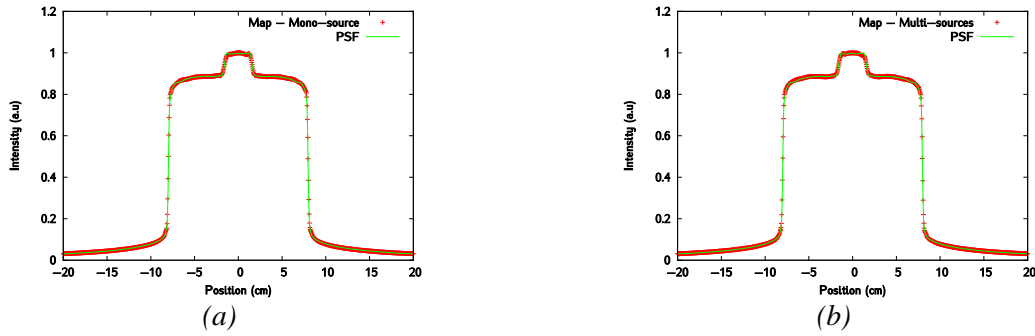


Figure 25. Heterogeneous phantom - Crossplane profiles of a  $10 \times 10 \text{ cm}^2$  field obtained with the reference PSF, the mono source models (a)  $\text{VSM}^{\text{MAP,mono}}$  and the multi-source models (b)  $\text{VSM}^{\text{MAP}}$ .

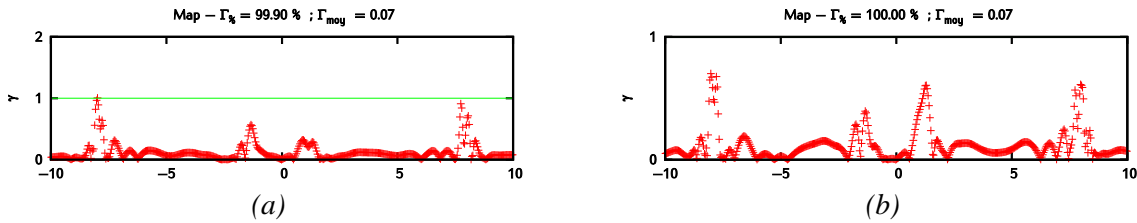


Figure 26. Heterogeneous phantom - Comparison of crossplane profiles of a  $10 \times 10 \text{ cm}^2$  field obtained with (a)  $\text{VSM}^{\text{MAP,mono}}$  and (b)  $\text{VSM}^{\text{MAP}}$  and reference PSF using a global 1% / 0 mm gamma index.

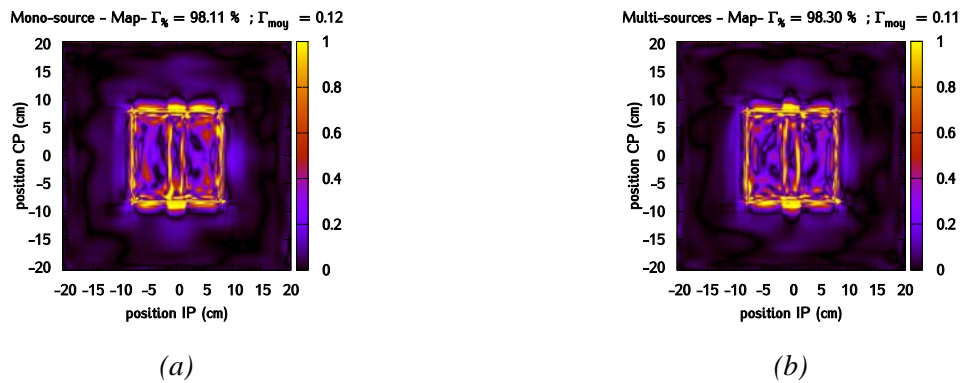


Figure 27. Heterogeneous phantom - Comparison of portal images of a  $10 \times 10 \text{ cm}^2$  field obtained with the VSM (a)  $\text{VSM}^{\text{MAP,mono}}$  and (b)  $\text{VSM}^{\text{MAP}}$  and reference PSF using a global 1% / 0 mm gamma index.

#### 4. Discussion

The aim of this study was to develop a generic VSM based on correlated histograms, usable for a wide range of applications in radiotherapy, which could necessitate different levels of dosimetric accuracy. In order to construct an accurate source model, we chose to keep all correlations existing between variables stored in the original PSF. This choice implies the handling of multi-dimensional histograms in a suitably chosen coordinate system: in our case, since four variables were used to characterize the particle behaviour, 4D-correlated histograms were used. In these 4D-histograms, the number of bins in each dimension (or for representing each variable), as well as their distribution, are key parameters of the source model, since they drive at the same time its accuracy (depending on its application), the statistical uncertainty of each bin, but also simulation runtime and random access memory issues. A trade-off has thereby to be found, and this observation led us to consider the development of adaptive binning schemes to construct an accurate but compact histogram-based VSM. To our knowledge, this issue was poorly investigated in previous studies, where authors have preferred to work with a regular

binning and to reduce the dimensionality of the problem by removing some correlations they considered as minor ones.

Here, a three-subsource VSM was developed and implemented in the PENELOPE MC code. Each sub-source was described by a 4D-histogram which stores the energy, the radial and the angular distributions generated from the original PSF. Four different adaptive binning schemes were investigated to map these distributions and construct the 4D-histograms, namely the *Quant*, *Cumul*, *ML* and *MAP* binnings. The performances of these four VSMs, namely  $VSM^{Quant}$ ,  $VSM^{Cumul}$ ,  $VSM^{ML}$  and  $VSM^{MAP}$ , were compared to a VSM constructed using a regular binning,  $VSM^{Reg}$ . Each of these VSM is stored on 300 Mo (100 Mo for each 4D-histogram) which is much more compact than the 71 Go PSF used to build them.

The performances of  $VSM^{Quant}$ ,  $VSM^{Cumul}$ ,  $VSM^{ML}$  and  $VSM^{MAP}$ , were first assessed in the phase-space plane, in order to verify with which accuracy the VSMs were able to reproduce the physical distributions from a 240 Go PSF (three time greater than the one used to build the histograms). At this step, the regular binning has already showed its inability to produce reliable dosimetric calculations. These bad results were confirmed by dose calculations in the water phantom, where outlying results were obtained. The regular binning was therefore not used in the following of the study. The *Reg* binning developed here leads to poorer results than those described in the literature. This was predictable: Unlike here, the authors of these studies do not take into account all the correlations existing between the variables. Histogram dimensions are lower than those built here. It is then possible to use more than 64 bins to describe each particle characteristic and then improve the model.

In the various dosimetric configurations that were simulated (water and heterogeneous phantoms, before and after collimation systems), the *Quant* binning gave also less good results for water tank calculations than the *Cumul* binning, and the *ML* and *MAP* binning methods based on the likelihood maximization. The *Cumul* binning is adaptive and simple to implement. Its limitation lies mainly in the quadratic criterion used for the construction which does not account for the histogram statistical properties as in *ML* and *MAP*. Only *MAP* binning scheme is however suitable to portal image prediction. A MSV dedicated to high-resolution portal image prediction has to be more accurate than those used for dosimetric calculations. The EPID pixel size is 5 to 6 times smaller than usual voxel size in dose distribution calculation in phantom or patient. Beside, errors due to MSV inaccuracies are amplified because the EPID is located at 160 cm from the target.

The impact of the adaptive binning schemes was also tested on a mono-source VSM, which was simply built without sorting out particles in the original PSF. Excellent dosimetric results were also obtained for this mono-source VSM, which enabled to compute 3D dose distributions and portal images as accurately as using the multi-source VSM, in the homogeneous phantom as well as in the heterogeneous phantom. This demonstrated that the use of a mono-source VSM built using a *MAP* binning method can be fully considered for 3D dosimetric as well as for portal image calculations, without compromising the accuracy. This is in contradiction with most of previously published studies claiming that multi-source VSMs outperform mono-source VSMs, which is true with binning strategies usually used. Indeed, the polar angle distributions in the VSM system are particularly distinct between the target and other sources. The great majority of the photons belonging to the T sub-source are not deviated after their emission by bremsstrahlung. This explains the peaked shape of the distribution around the central axis while for the other sub-sources secondary interactions contribute to enlarge it. These discrepancies cannot be handled by a single source with a regular or quantile binning. To the contrary an adaptive binning strategy considerably helps to capture irregularities in the distributions due to heterogeneous physical behaviours of particles. The interest of using a mono-source VSM instead of a multi-source one is obvious: its generation is easier whatever the modelled beam (including flattening filter free beams), and does not require a detailed analysis of the original PSF.

The interest of having more flexibility in the choice of the binning lies also in the fact that wider ranges of values for the different variables can be explored than with source models employing regular binning. To give an example, the VSM developed here allows representing the beam on a wide range of radial positions, going up to 15 cm radius at a SSD of 17 cm. This corresponds to maximum radial positions of about 1 m at the isocenter, which means that the VSM could potentially be used for out-of-field dose calculations, if appropriately commissioned. The adaptive binning hence provides the VSM developed here a higher flexibility which can be particularly valuable for applications in non-

standard conditions. Among them, a work is on-going to test the performances of these new VSMs with adaptive binning for the prediction of high-resolution EPID images (Lazaro et al., 2013).

Further refinements of our VSM could include the introduction of charged particles: this could be easily done by storing the characteristics of those particles in a separated PSF, which could be considered as an additional sub-source during the sampling procedure. This is for instance the choice retained by Fix (Fix et al., 2004). This step is mandatory especially when developing the VSM for higher linac energies. The integration of the collimation system modelling into the VSM would be undoubtedly the most important improvement to consider in future studies.

One shortcoming of our approach is the need to compute a reference PSF using MC simulation before constructing the VSM. This limitation is common to all histogram-based source models. Indeed, the reference must contain enough particles to guarantee that the statistical uncertainty in each bin is acceptable. From a statistical point of view, the simple empirical filling of the histogram can be seen as an *estimation* of the phase space distribution. As such, this estimate may reveal noisy in high dimensions even for huge statistics. Statistical methods involving regularity conditions in order to get smoother distributions seem, in our opinion, a valuable way to development, and should be explored.

## 5. Conclusions

This study presented the development of a generic VSM based on correlated histograms and its implementation in the MC code PENELOPE. This VSM was designed for a wide range of applications in radiotherapy which could differ by the degree of dosimetric accuracy that is required, portal image prediction being one of the most challenging. In order to preserve the highest degree of accuracy, we chose to design the VSM so that all correlations existing in the PSF from which it was derived could be kept. In this context, a special attention was then paid on the choice of the binning scheme used to construct the multi-dimensional histograms. This study highlighted the limitations of our VSM with a regular binning to describe with accuracy the energy, radial and angular distributions, and which led to erroneous dose calculations. The choice of an irregular and adaptive binning helped improving the dosimetric results, while balancing a trade-off between the size of the correlated histograms and the noise in each bin. Portal image calculation requires a sophisticated MSV such as the one based on the *MAP* method. Another important result obtained in this study was also the possibility to design a simple VSM with a single source describing the photons, regardless of their origin and history, at the condition of choosing a suited binning even in portal image prediction. The mono-source VSM constructed using the *MAP* binning demonstrated equivalent dosimetric performances than the three-source VSM constructed using the same binning. The advantage of using an adaptive binning during the elaboration of the VSM will be also a more important flexibility in the range of values for the representation of the different variables, allowing for instance to adapt the accuracy of the source model for out-of-field dose calculations or the computation of dose distributions with smaller dimensions of voxels. Future developments of the VSM will include the exploration of methods helping to control the level of noise in multi-dimensional histograms and offering the possibility to reduce the size of the reference PSF used to construct the model.

## Appendix A. From linac to VSM : Calculation of the polar and azimuthal angles

The purpose is here to convert the particle coordinate in the linac system into the VSM system. Each particle in the phase space file is characterized by fifteen variables as explained before (section 2.1). Six of them give its position and its direction:  $X, Y, Z, U, V, W$  where  $X, Y, Z$ , its Cartesian coordinates in the phase space plane,  $U, V, W$  the cosines of the direction of movement. Three variables in the VSM system,  $r_s, \varphi_{ab}, \theta_d$  correspond to these six variables in the linac system.

### A-1. Expression of the particle polar coordinates in the linac system

The first step is the calculation of the particle polar coordinates in the linac system. The particle radial position  $r_s$  is obtained using the equation (A- 1) :

$$r_s = \sqrt{X^2 + Y^2} \quad (\text{A- 1})$$

The angle  $\theta_{pos}$  corresponds to the particle radial position in the linac system:

$$\begin{cases} \theta_{pos} = \cos^{-1}\left(\frac{X}{r_s}\right) \text{ if } Y > 0 \\ \theta_{pos} = 2 * \pi * -\cos^{-1}\left(\frac{X}{r_s}\right) \text{ if } Y < 0 \end{cases} \quad (\text{A- 2})$$

The particle polar angle ,  $\varphi_{pos}$  is given by the expression:

$$\varphi_{pos} = \tan^{-1}\left(\frac{r_s}{Z}\right) \quad (\text{A- 3})$$

#### A-2. Expression of the direction in the VSM system

A new coordinate system is defined. The center of this system is the particle in the phase space plane. The (z') axis is the line linking the center of the target to the position of the particle in the phase space plane. The (x') axis is the line which is perpendicular to (z') and which crosses the original (z) axis.

The particle direction Cartesian coordinate is still given by U, V and W. A 3D-rotation is performed to express it in the VSM system. This 3D rotation is split in two 2D rotations. The first one consists in turning around the (z) of  $-\theta_{pos}$ .

$$\begin{pmatrix} U' \\ V' \\ W' \end{pmatrix} = \begin{pmatrix} \cos(\theta_{pos}) & -\sin(\theta_{pos}) & 0 \\ \sin(\theta_{pos}) & \cos(\theta_{pos}) & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} U \\ V \\ W \end{pmatrix} \quad (\text{A- 4})$$

The second one turns around the (y') of  $-\varphi_{pos}$ :

$$\begin{pmatrix} U'' \\ V'' \\ W'' \end{pmatrix} = \begin{pmatrix} \cos(\varphi_{pos}) & 0 & -\sin(\varphi_{pos}) \\ 0 & 1 & 0 \\ \sin(\varphi_{pos}) & 0 & \cos(\varphi_{pos}) \end{pmatrix} \cdot \begin{pmatrix} U' \\ V' \\ W' \end{pmatrix} \quad (\text{A- 5})$$

The vector (U'', V'', W'') is the particle direction in Cartesian coordinates in the VSM system.

#### A-3. Expression of the direction in polar coordinates

The last step is the conversion of the vector (U'', V'', W'') in polar coordinates:

$$r_{new} = \sqrt{U''^2 + W''^2} \quad (\text{A- 6})$$

$$\alpha_{new} = \cos^{-1}\left(\frac{U''}{r_{new}}\right) \quad (\text{A- 7})$$

$$\begin{cases} \theta_{d_{rad}} = \alpha_{new} \text{ if } V'' \geq 0 \\ \theta_{d_{rad}} = 2 * \pi - \alpha_{new} \text{ if } V'' < 0 \end{cases} \quad (\text{A- 8})$$

$$\begin{cases} \theta_{d'_{rad}} = \theta_{d_{rad}} - 2 * \pi \text{ if } \theta_{d_{rad}} > \pi \\ \theta_{d'_{rad}} = \theta_{d_{rad}} \\ \theta_{d'_{rad}} = \theta_{d_{rad}} + 2 * \pi \text{ if } \theta_{d_{rad}} < -\pi \end{cases} \quad (\text{A- 9})$$

$$\theta_d = \frac{\theta_{d_{rad}} * 180}{\pi} \quad (\text{A- 10})$$

$$\varphi_d = \cos^{-1} \left( \frac{W''}{\pi} * 180 \right) \quad (\text{A- 11})$$

The particle is finally described by three geometric variables:  $r_s$ ,  $\varphi_d$ , and  $\theta_d$ .

### Appendix B. From VSM to linac : Calculation of the polar and azimuthal angles

A particle generated with the VSM is described with four variables:  $E$ ,  $r_s$ ,  $\varphi_d$ , and  $\theta_d$ . Our goal is this time to convert them in the linac system and express them as  $X$ ,  $Y$ ,  $Z$ ,  $U$ ,  $V$  and  $W$ .

A random number 'a' is sampled between 0 and 1.

$$\theta_r = a * 2 * \pi \quad (\text{B- 1})$$

The particle position in the linac system is then:

$$\begin{cases} X = r_s * \cos(\theta_r) \\ Y = r_s * \sin(\theta_r) \\ Z = z_{VSM} \end{cases} \quad (\text{B- 2})$$

The polar and azimuthal angles  $\varphi_d$  and  $\theta_d$  given the particle direction in the VSM system are first converted in radian, and then expressed in Cartesian coordinates:

$$\theta_{d_{rad}} = \frac{\theta_d * \pi}{180} \quad (\text{B- 3})$$

$$\varphi_{d_{rad}} = \frac{\varphi_d * \pi}{180} \quad (\text{B- 4})$$

$$\begin{cases} dx = \cos(\theta_d) * \sin(\varphi_d) \\ dy = \sin(\theta_d) * \sin(\varphi_d) \\ dz = \cos(\varphi_d) \end{cases} \quad (\text{B- 5})$$

A first rotation of  $\varphi'_d$  around the (y') axis is done, and then a second one of  $\theta_r$  around the (z').

$$\varphi'_d = \tan^{-1} \left( \frac{r'_s}{Z} \right) \quad (\text{B- 6})$$

$$\begin{pmatrix} U' \\ V' \\ W' \end{pmatrix} = \begin{pmatrix} \cos(\varphi'_d) & 0 & \sin(\varphi'_d) \\ 0 & 1 & 0 \\ -\sin(\varphi'_d) & 0 & \cos(\varphi'_d) \end{pmatrix} \cdot \begin{pmatrix} dx \\ dy \\ dz \end{pmatrix} \quad (\text{B- 7})$$

$$\begin{pmatrix} U \\ V \\ W \end{pmatrix} = \begin{pmatrix} \cos(\theta_r) & -\sin(\theta_r) & 0 \\ \sin(\theta_r) & \cos(\theta_r) & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} U' \\ V' \\ W' \end{pmatrix} \quad (\text{B- 8})$$

The particle characteristics are now expressed under the PENELOPE form  $X$ ,  $Y$ ,  $Z$ ,  $U$ ,  $V$ ,  $W$ .

### Appendix C. Binning optimization and likelihood maximization penalized likelihood maximization.

#### C-1. Maximum likelihood approach

In this approach our aim is to adjust the width of histograms bins using the Maximum Likelihood criterion (leading to the *ML* source). This optimization is considered separately for each dimension of the source model.

The likelihood  $L_I$  is in the case of irregular histograms defined by the equation (C- 1):

$$L_1 = \log(P(X|I)) \quad (\text{C- 1})$$

where  $X$  is the random repartition of PSF particles over intervals  $I$ . This partition is constituted of  $D$  segments whose  $j^{\text{th}}$  width is denoted by  $I_j$ . In our study we used  $D = 64$ . If  $n$  and  $N_j$  are respectively the total number of particles and the particles number in bin  $j$ , then  $X = \{N_1, N_2, \dots, N_D\}$ ,  $I = \{I_1, I_2, \dots, I_D\}$  and  $P(X|I)$  is written :

$$P(X|I) = \left( \prod_{j=1}^D \left( \frac{N_j}{n|I_j|} \right)^{N_j} \right) \quad (\text{C- 2})$$

This expression corresponds to the VSM model where the density is uniform in each bin. The probability density function of the empirical histogram associated to the partition  $I$  is  $\frac{N_j}{n|I_j|}$  for bin  $j$ .

The probability to observe  $N_j$  independent particles in this bin is thus  $\left( \frac{N_j}{n|I_j|} \right)^{N_j}$ . Since the  $D$  bins are assumed independent, we derive the expression for  $P(X|I)$ .

This leads to the expression of the log-likelihood :

$$L_1 = \sum_{j=1}^D N_j \cdot \log \left( \frac{N_j}{n|I_j|} \right) \quad (\text{C- 3})$$

We can interpret  $(-L_I)$  as a cost function which has to be minimized. We then use a quantitative criterion for approximating the phase state space by a histogram. The *ML* binning is thus optimal with respect of this criterion.

The following algorithm which solves the  $(-L_I)$  minimization problem is based on a linear programming method (Rozenholc et al., 2010). We can distinguish three main steps:

- (1) First  $I$  is segmented in 2048 regular micro-bins.
- (2) All these 2048 micro-bins are combined together following all possible contiguous combinations in order to produce 64 macro-bins.
- (3) The best association, aka the best partition  $I$ , corresponds to the maximization of  $L_I$ . We retain this solution.

#### C-2. Penalization of ML solution through quantiles regularity : maximum *a posteriori* (leading to the *MAP* source)

This modification of maximum likelihood method tends to give more regularity in the bins widths by considering a *prior* on these widths encouraging bins of equal areas. The interest of this regularization appears when bins are rather noisy (working with moderate statistics PSF). In this case, the *ML* solution may reveal irregular bins widths. Even if this tendency vanishes with huge PSF, we propose a general approach robust to moderate statistics conditions. This binning strategy is again inspired by



Rozenholc *et al.* (Rozenholc et al., 2010) and can be viewed as the combination of maximum likelihood and quantiles methods.

We used a symmetric Dirichlet distribution for the prior of bins area. Using Bayes theorem, with same notations as above we can write

$$P(X, I) = P(I|X)P(X) = P(X|I)P(I) \quad (\text{C- 4})$$

Leading to the penalized log-likelihood  $L_2$  :

$$L_2 = \log(P(X|I).P(I)) \quad (\text{C- 5})$$

where  $P(I)$  stands for the Dirichlet distribution ::

$$P(I) = K. \left(\frac{N_1}{n}\right)^{\alpha-1} \cdot \left(\frac{N_2}{n}\right)^{\alpha-1} \dots \left(\frac{N_D}{n}\right)^{\alpha-1} \quad (\text{C- 6})$$

Where  $\alpha$  is a positive tuning parameter and  $K$  is a normalizing constant. The expression of the penalized log-likelihood becomes :

$$L_2 = \sum_{j=1}^D N_j. \log\left(\frac{N_j}{n. |I_j|}\right) + (\alpha - 1). \sum_{j=1}^D \log\left(\frac{N_j}{n}\right) \quad (\text{C- 7})$$

Particular choices of  $\alpha$  allows to mitigate the binning between the *ML* solution and the quantile prior. When  $\alpha = 0$  , we recover the *ML* source. When  $\alpha$  is large we tend to the quantile binning. In our study, we found that  $\alpha$  values in the order of 100 are convenient. Indeed, following heuristic considerations,  $L_2$  expression suggests that  $\alpha$  can be roughly interpreted as a number of virtual particles. We can thus approximately consider that for bins with less than  $\alpha$  particles, the *prior* influence is significant. The  $L_2$  maximization follows the same linear programming scheme as for  $L_1$  maximization.

## References

- Ahnesjö, A., Weber, L., Murman, A., Saxner, M., Thorslund, I., Traneus, E., 2005. Beam modeling and verification of a photon beam multisource model. *Med. Phys.* 32, 1722–1737.
- Badel, J.-N., 2009. Contrôle dosimétrique des traitements de radiothérapie par simulation Monte Carlo de l'image de dose portale transmise. L'Institut National des Sciences Appliquées de Lyon.
- Cashmore, J., Golubev, S., Dumont, J.L., Sikora, M., Alber, M., Ramtohul, M., 2012. Validation of a virtual source model for Monte Carlo dose calculations of a flattening filter free linac. *Med. Phys.* 39, 3262–3269. doi:10.1118/1.4709601
- Deng, J., Jiang, S.B., Kapur, A., Li, J., Pawlicki, T., Ma, C.M., 2000. Photon beam characterization and modelling for Monte Carlo treatment planning. *Phys. Med. Biol.* 45, 411–427.
- Devroye, L., 1986. Non-Uniform Random Variate Generation.
- Faddegon, B.A., Blevis, I., 2000. Electron spectra derived from depth dose distributions. *Med. Phys.* 27, 514–526.
- Fippel, M., Haryanto, F., Dohm, O., Nusslin, F., Kriesen, S., 2003. A virtual photon energy fluence model for Monte Carlo dose calculation. *Med. Phys.* 30, 301–311. doi:10.1118/1.1543152
- Fix, M., Keall, P., Libby, B., Mohan, R., Siebers, J., 2004. Monte Carlo source model for megavoltage photon-beam radiotherapy: Photon source characteristics. *Med. Phys.* 31, 1904–1904.
- Fix, M.K., Keller, H., Ruegsegger, P., Born, E.J., 2000. Simple beam models for Monte Carlo photon beam dose calculations in radiotherapy. *Med. Phys.* 27, 2739–2747. doi:10.1118/1.1318220
- Fix, M.K., Manser, P., Born, E.J., Mini, R., Ruegsegger, P., 2001a. Monte Carlo simulation of a dynamic MLC based on a multiple source model. *Phys. Med. Biol.* 46, 3241–3257.
- Fix, M.K., Stampanoni, M., Manser, P., Born, E.J., Mini, R., Ruegsegger, P., 2001b. A multiple source model for 6 MV photon beam dose calculations using Monte Carlo. *Phys. Med. Biol.* 46, 1407–1427.

- Janssen, J.J., Korevaar, E.W., Battum, L.J. van, Storchi, P.R.M., Huizenga, H., 2001. A model to determine the initial phase space of a clinical electron beam from measured beam data. *Phys. Med. Biol.* 46, 269–286. doi:10.1088/0031-9155/46/2/301
- Lazaro, D., Barat, E., Loirec, C.L., Dautremer, T., Montagu, T., Guérin, L., Batalla, A., 2013. Denoising techniques combined to Monte Carlo simulations for the prediction of high-resolution portal images in radiotherapy treatment verification. *Phys. Med. Biol.* 58, 3433. doi:10.1088/0031-9155/58/10/3433
- Ma, C.M., Faddegon, B.A., Rogers, D.W., Mackie, T.R., 1997. Accurate characterization of Monte Carlo calculated electron beams for radiotherapy. *Med. Phys.* 24, 401–416.
- Rozenholc, Y., Mildenerger, T., Gather, U., 2010. Combining Regular and Irregular Histograms by Penalized Likelihood. *Comput Stat Data Anal* 54, 3313–3323. doi:10.1016/j.csda.2010.04.021
- Schach von Wittenau, A.E., Cox, L.J., Bergstrom, P.M., Jr, Chandler, W.P., Hartmann Siantar, C.L., Mohan, R., 1999. Correlated histogram representation of Monte Carlo derived medical accelerator photon-output phase space. *Med. Phys.* 26, 1196–1211.
- Sempau, J., Sánchez-Reyes, A., Salvat, F., ben Tahar, H.O., Jiang, S.B., Fernández-Varea, J.M., 2001. Monte Carlo simulation of electron beams from an accelerator head using PENELOPE. *Phys. Med. Biol.* 46, 1163–1186.
- Sikora, M., Dohm, O., Alber, M., 2007. A virtual photon source model of an Elekta linear accelerator with integrated mini MLC for Monte Carlo based IMRT dose calculation. *Phys. Med. Biol.* 52, 4449–4463. doi:10.1088/0031-9155/52/15/006
- Tillikainen, L., Siljamäki, S., 2008. A multiple-source photon beam model and its commissioning process for VMC++ Monte Carlo code. *J. Phys. Conf. Ser.* 102, 012024. doi:10.1088/1742-6596/102/1/012024
- Wilding, G.E., Mudholkar, G.S., 2008. Empirical approximations for Hoeffding's test of bivariate independence using two Weibull extensions. *Stat. Methodol.* 5, 160–170. doi:10.1016/j.stamet.2007.07.002