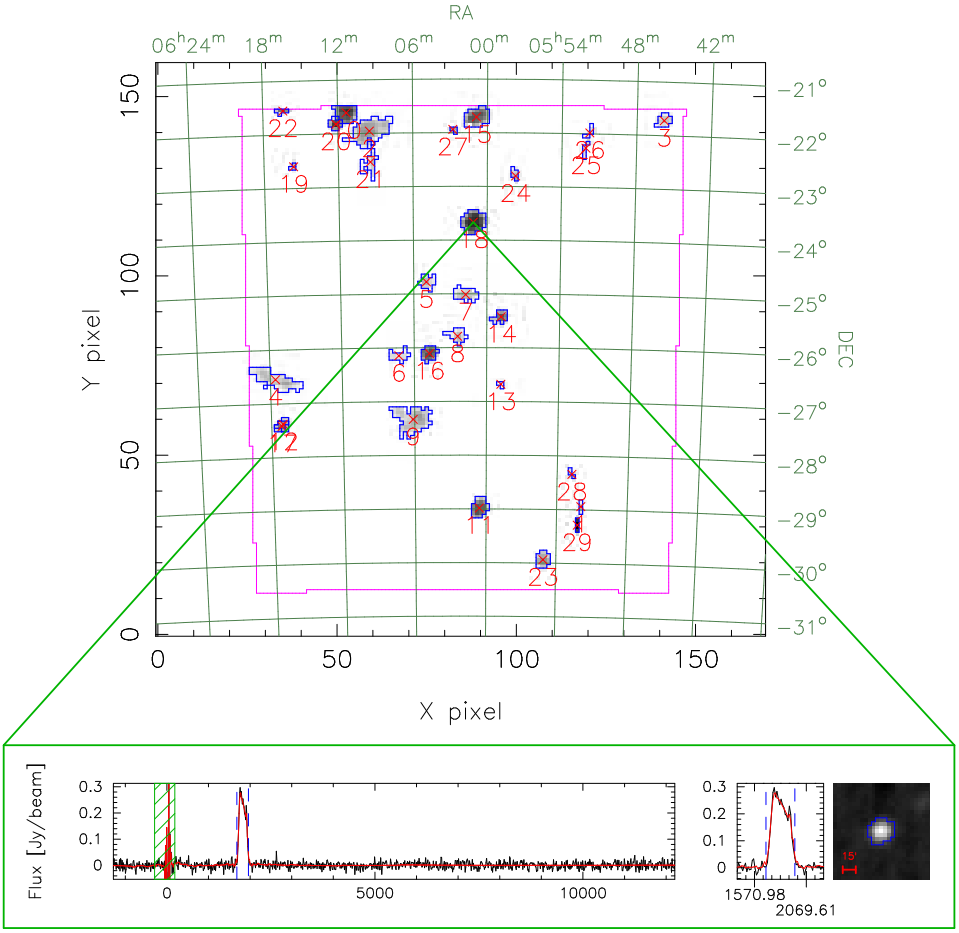
Source Detection with *Duchamp*

A User's Guide

Matthew Whiting
Australia Telescope National Facility
CSIRO

Duchamp version 1.1
May 17, 2007



Contents

1 Introduction and getting going quickly **4**
1.1 About Duchamp 4
1.2 What to do 4
1.3 Guide to terminology and conventions 5
1.4 A summary of the execution steps 6

2 User Inputs **9**

3 What *Duchamp* is doing **10**
3.1 Image input 10
3.2 Image modification 11
3.2.1 BLANK pixel removal 11
3.2.2 Baseline removal 11
3.2.3 Ignoring bright Milky Way emission 12
3.3 Image reconstruction 12
3.3.1 Algorithm 12
3.3.2 Note on Statistics 13
3.3.3 User control of reconstruction parameters 14
3.4 Smoothing the cube 14
3.4.1 Spectral smoothing 15
3.4.2 Spatial smoothing 15
3.5 Input/Output of reconstructed/smoothed arrays 16
3.6 Searching the image 17
3.6.1 Technique 17
3.6.2 Calculating statistics 17
3.6.3 Determining the threshold 18
3.7 Merging and growing detected objects 19

4 Outputs **21**
4.1 During execution 21
4.2 Text-based output files 22
4.2.1 Table of results 22
4.2.2 Other results lists 24
4.3 Graphical output 25
4.3.1 Spectral plots 25
4.3.2 Spatial maps 26

5 Notes and hints on the use of *Duchamp* **28**

6 Future developments **30**

7 Why "*Duchamp*"? **30**

A Obtaining and installing *Duchamp* **32**
A.1 Installing 32
A.2 Running *Duchamp* 33
A.3 Feedback 33
A.4 Beta Versions 34

B Available parameters **35**

C Example parameter files **41**

D Example results file **43**

E Example VOTable output **45**

F Example Karma Annotation file output **46**

G Robust statistics for a Normal distribution **47**

H How Gaussian noise changes with wavelet scale **49**

1 Introduction and getting going quickly

1.1 About Duchamp

This document provides a user's guide to *Duchamp*, an object-finder for use on spectral-line data cubes. The basic execution of *Duchamp* is to read in a FITS data cube, find sources in the cube, and produce a text file of positions, velocities and fluxes of the detections, as well as a postscript file of the spectra of each detection.

Duchamp has been designed to search for objects in particular sorts of data: those with relatively small, isolated objects in a large amount of background or noise. Examples of such data are extragalactic HI surveys, or maser surveys. *Duchamp* searches for groups of connected voxels (or pixels) that are all above some flux threshold. No assumption is made as to the shape of detections, and the only size constraints applied are those specified by the user.

1.2 What to do

So, you have a FITS cube, and you want to find the sources in it. What do you do? First, you need to get *Duchamp*: there are instructions in Appendix A for obtaining and installing it. Once you have it running, the first step is to make an input file that contains the list of parameters. Brief and detailed examples are shown in Appendix C. This file provides the input file name, the various output files, and defines various parameters that control the execution.

The standard way to run *Duchamp* is by the command

```
> Duchamp -p [parameter file]
```

replacing `[parameter file]` with the name of the file listing the parameters.

An even easier way is to use the default values for all parameters (these are given in Appendix B and in the file `InputComplete` included in the distribution directory) and use the syntax

```
> Duchamp -f [FITS file]
```

where `[FITS file]` is the file you wish to search.

The default action includes displaying a map of detected objects in a PGPLOT X-window. This can be disabled by setting the parameter `flagXOutput = false` or using the `-x` command-line option, as in

```
> Duchamp -x -p [parameter file]
```

and similarly for the `-f` case.

Once a FITS file and parameters have been set, the program will then work away and give you the list of detections and their spectra. The program execution is summarised below, and detailed in §3. Information on inputs is in §2 and Appendix B, and descriptions of the output is in §4.

1.3 Guide to terminology and conventions

First, a brief note on the use of terminology in this guide. *Duchamp* is designed to work on FITS "cubes". These are FITS[1] image arrays with (at least) three dimensions. They are assumed to have the following form: the first two dimensions (referred to as x and y) are spatial directions (that is, relating to the position on the sky – often, but not necessarily, corresponding to Equatorial or Galactic coordinates), while the third dimension, z, is the spectral direction, which can correspond to frequency, wavelength, or velocity. The three dimensional analogue of pixels are "voxels", or volume cells – a voxel is defined by a unique (x, y, z) location and has a single value of flux, intensity or brightness (or something equivalent) associated with it.

Sometimes, some pixels in a FITS file are labelled as BLANK – that is, they are given a nominal value, defined by FITS header keywords BLANK, BSCALE, & BZERO, that marks them as not having a flux value. These are often used to pad a cube out so that it has a rectangular spatial shape. *Duchamp* has the ability to avoid these: see §3.2.1.

Note that it is possible for the FITS file to have more than three dimensions (for instance, there could be a fourth dimension representing a Stokes parameter). Only the two spatial dimensions and the spectral dimension are read into the array of pixel values that is searched for objects. All other dimensions are ignored[2]. Herein, we discuss the data in terms of the three basic dimensions, but you should be aware it is possible for the FITS file to have more than three. Note that the order of the dimensions in the FITS file does not matter.

With this setup, each spatial pixel (a given (x, y) coordinate) can be said to be a single spectrum, while a slice through the cube perpendicular to the spectral direction at a given z-value is a single channel, with the 2-D image in that channel called a channel map.

Detection involves locating a contiguous group of voxels with fluxes above a certain threshold. *Duchamp* makes no assumptions as to the size or shape of the detected features, other than having user-selected minimum size criteria. Features that are detected are assumed to be positive. The user can choose to search for negative features by setting an input parameter – this inverts the cube prior to the search (see §3.6 for details).

[1] FITS is the Flexible Image Transport System – see Hanisch et al. (2001) or websites such as http://fits.cv.nrao.edu/FITS.html for details.

[2] This actually means that the first pixel only of that axis is used, and the array is read by the `fits_read_subsetnull` command from the CFITSIO library.

Finally, note that it is possible to run *Duchamp* on a two-dimensional image (i.e. one with no frequency or velocity information), or indeed a one-dimensional array, and many of the features of the program will work fine. The focus, however, is on object detection in three dimensions, one of which is a spectral dimension.

1.4 A summary of the execution steps

The basic flow of the program is summarised here – all steps are discussed in more detail in the following sections.

1. The necessary parameters are recorded.

   How this is done depends on the way the program is run from the command line. If the `-p` option is used, the parameter file given on the command line is read in, and the parameters therein are read. All other parameters are given their default values (listed in Appendix B).

   If the `-f` option is used, all parameters are assigned their default values.

2. The FITS image is located and read in to memory.

   The file given is assumed to be a valid FITS file. As discussed above, it can have any number of dimensions, but *Duchamp* only reads in the two spatial and the one spectral dimensions. A subset of the FITS array can be given (see §3.1 for details).

3. If requested, a FITS file containing a previously reconstructed or smoothed array is read in.

   When a cube is either smoothed or reconstructed with the *à trous* wavelet method, the result can be saved to a FITS file, so that subsequent runs of *Duchamp* can read it in to save having to re-do the calculations (as they can be relatively time-intensive).

4. If requested, BLANK pixels are trimmed from the edges, and the baseline of each spectrum is removed.

   BLANK pixels, while they are ignored by all calculations in *Duchamp*, do increase the size in memory of the array above that absolutely needed. This step trims them from the spatial edges, recording the amount trimmed so that they can be added back in later.

   A spectral baseline (or bandpass) can also be removed at this point as well. This may be necessary if there is a ripple or other large-scale feature present that will hinder detection of faint sources.

5. If the reconstruction method is requested, and the reconstructed array has not been read in at Step 3 above, the cube is reconstructed using the *à trous* wavelet method.

This step uses the *à trous* method to determine the amount of structure present at various scales. A simple thresholding technique then removes random noise from the cube, leaving the significant signal. This process can greatly reduce the noise level in the cube, enhancing the detectability of sources.

6. Alternatively (and if requested), the cube is smoothed, either spectrally or spatially.

   This step presents two options. The first considers each spectrum individually, and convolves it with a Hanning filter (with width chosen by the user). The second considers each channel map separately, and smoothes it with a Gaussian kernel of size and shape chosen by the user. This step can help to reduce the amount of noise visible in the cube and enhance fainter sources.

7. A threshold for the cube is then calculated, based on the pixel statistics (unless a threshold is manually specified by the user).

   The threshold can either be chosen as a simple $n\sigma$ threshold (i.e. a certain number of standard deviations above the mean), or calculated via the "False Discovery Rate" method. Alternatively, the threshold can be specified as a simple flux value, without care as to the statistical significance (e.g. "I want every source brighter than 10mJy").

   By default, the full cube is used for the statistics calculation, although the user can nominate a subsection of the cube to be used instead.

8. Searching for objects then takes place, using the requested thresholding method.

   The cube is searched one channel-map at a time. Detections are compared to already detected objects and either combined with a neighbouring one or added to the end of the list.

9. The list of objects is condensed by merging neighbouring objects and removing those deemed unacceptable.

   While some merging has been done in the previous step, this process is a much more rigorous comparison of each object with every other one. If a pair of objects lie within requested limits, they are combined.

   After the merging is done, the list is culled (although see comment for the next step). There are certain criteria the user can specify that objects must meet: minimum numbers of spatial pixels and spectral channels, and minimum separations between neighbouring objects. Those that do not meet these criteria are deleted from the list.

10. If requested, the objects are "grown" down to a lower threshold, and then the merging step is done a second time.

In this case, each object has pixels in its neighbourhood examined, and if they are above a secondary threshold, they are added to the object. The merging process is done a second time in case two objects have grown over the top of one another. Note that the rejection part of the previous step is not done until the end of the second merging process.

11. The baselines and trimmed pixels are replaced prior to output.

    This is just the inverse of step #4.

12. The details of the detections are written to screen and to the requested output file.

    Crucial properties of each detection are provided, showing its location, extent, and flux. These are presented in both pixel coordinates and world coordinates (e.g. sky position and velocity). Any warning flags are also printed, showing detections to be wary of. Alternative output options are available, such as a VOTable or a Karma annotation file.

13. Maps showing the spatial location of the detections are written.

    These are 2-dimensional maps, showing where each detection lies on the spatial coverage of the cube. This is provided as an aid to the user so that a quick idea of the distribution of object positions can be gained e.g. are all the detections on the edge?

    Two maps are provided: one is a 0th moment map, showing the 0th moment (i.e. a map of the integrated flux) of each detection in its appropriate position, while the second is a "detection map", showing the number of times each spatial pixel was detected in the searching routines (including those pixels rejected at step 9 and so not in any of the final detections).

    These maps are written to postscript files, and the 0th moment map can also be displayed in a PGPLOT X-window.

14. The integrated or peak spectra of each detection are written to a postscript file.

    The spectral equivalent of the maps – what is the spectral profile of each detection? Also provided here are basic information for each object (a summary of the information in the results file), as well as a 0th moment map of the detection.

15. If requested, the reconstructed or smoothed array can be written to a new FITS file.

    If either of these procedures were done, the resulting array can be saved as a FITS file for later use. The FITS header will be the same as the input file, with a few additional keywords to identify the file.

2 User Inputs

Input to the program is provided by means of a parameter file. Parameters are listed in the file, followed by the value that should be assigned to them. The syntax used is

```
parameterName value.
```

Parameter names are not case-sensitive, and lines in the input file that start with `#` are ignored. If a parameter is listed more than once, the latter value is used, but otherwise the order in which the parameters are listed in the input file is arbitrary. Example input files can be seen in Appendix C.

If a parameter is not listed, the default value is assumed. The defaults are chosen to provide a good result (using the reconstruction method), so the user doesn't need to specify many new parameters in the input file. Note that the image file **must** be specified! The parameters that can be set are listed in Appendix B, with their default values in parentheses.

The parameters with names starting with `flag` are stored as `bool` variables, and so are either `true = 1` or `false = 0`. They can be entered in the file either in text or integer format – *Duchamp* will read them correctly in either case.

An example input file is included in the distribution tar file. It is as follows:

```
imageFile       your-file-here
logFile         logfile.txt
outFile         results.txt
spectraFile     spectra.ps
minPix          2
flagATrous      1
snrRecon        5.
snrCut          3.
minChannels     3
flagBaseline    1
```

You would, of course, replace the "`your-file-here`" with the FITS file you wanted to search. Further examples are given in Appendix C.

3 What *Duchamp* is doing

Each of the steps that *Duchamp* goes through in the course of its execution are discussed here in more detail. This should provide enough background information to fully understand what *Duchamp* is doing and what all the output information is. For those interested in the programming side of things, *Duchamp* is written in C/C++ and makes use of the CFITSIO, WCSLIB and PGPLOT libraries.

3.1 Image input

The cube is read in using basic CFITSIO commands, and stored as an array in a special C++ class. This class keeps track of the list of detected objects, as well as any reconstructed arrays that are made (see §3.3). The World Coordinate System (WCS)[3] information for the cube is also obtained from the FITS header by WCSLIB functions (Calabretta & Greisen 2002; Greisen & Calabretta 2002), and this information, in the form of a `wcsprm` structure, is also stored in the same class.

A sub-section of a cube can be requested by defining the subsection with the `subsection` parameter and setting `flagSubsection = true` – this can be a good idea if the cube has very noisy edges, which may produce many spurious detections.

There are two ways of specifying the `subsection` string. The first is the generalised form `[x1:x2:dx,y1:y2:dy,z1:z2:dz,...]`, as used by the CFITSIO library. This has one set of colon-separated numbers for each axis in the FITS file. In this manner, the x-coordinates run from `x1` to `x2` (inclusive), with steps of `dx`. The step value can be omitted, so a subsection of the form `[2:50,2:50,10:1000]` is still valid. In fact, *Duchamp* does not make use of any step value present in the subsection string, and any that are present are removed before the file is opened.

If the entire range of a coordinate is required, one can replace the range with a single asterisk, e.g. `[2:50,2:50,*]`. Thus, the subsection string `[*,*,*]` is simply the entire cube. Note that the pixel ranges for each axis start at 1, so the full pixel range of a 100-pixel axis would be expressed as 1:100. A complete description of this section syntax can be found at the FITSIO web site[4].

Making full use of the subsection requires knowledge of the size of each of the dimensions. If one wants to, for instance, trim a certain number of pixels off the edges of the cube, without examining the cube to obtain the actual size, one can use the second form of the subsection string. This just

[3] This is the information necessary for translating the pixel locations to quantities such as position on the sky, frequency, velocity, and so on.

[4] http://heasarc.gsfc.nasa.gov/docs/software/fitsio/c/c_user/node91.html

gives a number for each axis, e.g. `[5,5,5]` (which would trim 5 pixels from the start *and* end of each axis).

All types of subsections can be combined e.g. `[5,2:98,*]`.

3.2 Image modification

Several modifications to the cube can be made that improve the execution and efficiency of *Duchamp* (their use is optional, governed by the relevant flags in the parameter file).

3.2.1 BLANK pixel removal

If the imaged area of a cube is non-rectangular (see the example in Fig. 2, a cube from the HIPASS survey), BLANK pixels are used to pad it out to a rectangular shape. The value of these pixels is given by the FITS header keywords BLANK, BSCALE and BZERO. While these pixels make the image a nice shape, they will take up unnecessary space in memory, and so to potentially speed up the processing we can trim them from the edge. This is done when the parameter `flagTrim = true`. If the above keywords are not present, the trimming will not be done (in this case, a similar effect can be accomplished, if one knows where the "blank" pixels are, by using the subsection option).

The amount of trimming is recorded, and these pixels are added back in once the source-detection is completed (so that quoted pixel positions are applicable to the original cube). Rows and columns are trimmed one at a time until the first non-BLANK pixel is reached, so that the image remains rectangular. In practice, this means that there will be some BLANK pixels left in the trimmed image (if the non-BLANK region is non-rectangular). However, these are ignored in all further calculations done on the cube.

3.2.2 Baseline removal

Second, the user may request the removal of baselines from the spectra, via the parameter `flagBaseline`. This may be necessary if there is a strong baseline ripple present, which can result in spurious detections at the high points of the ripple. The baseline is calculated from a wavelet reconstruction procedure (see §3.3) that keeps only the two largest scales. This is done separately for each spatial pixel (i.e. for each spectrum in the cube), and the baselines are stored and added back in before any output is done. In this way the quoted fluxes and displayed spectra are as one would see from the input cube itself – even though the detection (and reconstruction if applicable) is done on the baseline-removed cube.

The presence of very strong signals (for instance, masers at several hundred Jy) could affect the determination of the baseline, and would lead to a large dip centred on the signal in the baseline-subtracted spectrum. To

prevent this, the signal is trimmed prior to the reconstruction process at some standard threshold (at 8σ above the mean). The baseline determined should thus be representative of the true, signal-free baseline. Note that this trimming is only a temporary measure which does not affect the source-detection.

3.2.3 Ignoring bright Milky Way emission

Finally, a single set of contiguous channels can be ignored – these may exhibit very strong emission, such as that from the Milky Way as seen in extragalactic HI cubes (hence the references to "Milky Way" in relation to this task – apologies to Galactic astronomers!). Such dominant channels will produce many detections that are unnecessary, uninteresting (if one is interested in extragalactic HI) and large (in size and hence in memory usage), and so will slow the program down and detract from the interesting detections.

The use of this feature is controlled by the `flagMW` parameter, and the exact channels concerned are able to be set by the user (using `maxMW` and `minMW` – these give an inclusive range of channels). When employed, these channels are ignored for the searching, and the scaling of the spectral output (see Fig. 1) will not take them into account. They will be present in the reconstructed array, however, and so will be included in the saved FITS file (see §3.5). When the final spectra are plotted, the range of channels covered by these parameters is indicated by a green hashed box.

3.3 Image reconstruction

The user can direct *Duchamp* to reconstruct the data cube using the *à trous* wavelet procedure. A good description of the procedure can be found in Starck & Murtagh (2002). The reconstruction is an effective way of removing a lot of the noise in the image, allowing one to search reliably to fainter levels, and reducing the number of spurious detections. This is an optional step, but one that greatly enhances the source-detection process, with the payoff that it can be relatively time- and memory-intensive.

3.3.1 Algorithm

The steps in the *à trous* reconstruction are as follows:

1. The reconstructed array is set to 0 everywhere.
2. The input array is discretely convolved with a given filter function. This is determined from the parameter file via the `filterCode` parameter – see Appendix B for details on the filters available.

3. The wavelet coefficients are calculated by taking the difference between the convolved array and the input array.

4. If the wavelet coefficients at a given point are above the requested threshold (given by `snrRecon` as the number of σ above the mean and adjusted to the current scale – see Appendix H), add these to the reconstructed array.

5. The separation between the filter coefficients is doubled. (Note that this step provides the name of the procedure[5], as gaps or holes are created in the filter coverage.)

6. The procedure is repeated from step 2, using the convolved array as the input array.

7. Continue until the required maximum number of scales is reached.

8. Add the final smoothed (i.e. convolved) array to the reconstructed array. This provides the "DC offset", as each of the wavelet coefficient arrays will have zero mean.

The reconstruction has at least two iterations. The first iteration makes a first pass at the wavelet reconstruction (the process outlined in the 8 stages above), but the residual array will likely have some structure still in it, so the wavelet filtering is done on the residual, and any significant wavelet terms are added to the final reconstruction. This step is repeated until the change in the measured standard deviation of the background (see note below on the evaluation of this quantity) is less than some fiducial amount.

It is important to note that the *à trous* decomposition is an example of a "redundant" transformation. If no thresholding is performed, the sum of all the wavelet coefficient arrays and the final smoothed array is identical to the input array. The thresholding thus removes only the unwanted structure in the array.

Note that any BLANK pixels that are still in the cube will not be altered by the reconstruction – they will be left as BLANK so that the shape of the valid part of the cube is preserved.

3.3.2 Note on Statistics

The correct calculation of the reconstructed array needs good estimators of the underlying mean and standard deviation of the background noise distribution. These statistics are estimated using robust methods, to avoid corruption by strong outlying points. The mean of the distribution is actually estimated by the median, while the median absolute deviation from the median (MADFM) is calculated and corrected assuming Gaussianity to

[5] *à trous* means "with holes" in French.

estimate the underlying standard deviation σ. The Gaussianity (or Normality) assumption is critical, as the MADFM does not give the same value as the usual rms or standard deviation value – for a Normal distribution $N(\mu, \sigma)$ we find MADFM= 0.6744888σ, but this will change for different distributions. Since this ratio is corrected for, the user need only think in the usual multiples of the rms when setting `snrRecon`. See Appendix G for a derivation of this value.

When thresholding the different wavelet scales, the value of the rms as measured from the wavelet array needs to be scaled to account for the increased amount of correlation between neighbouring pixels (due to the convolution). See Appendix H for details on this scaling.

3.3.3 User control of reconstruction parameters

The most important parameter for the user to select in relation to the reconstruction is the threshold for each wavelet array. This is set using the `snrRecon` parameter, and is given as a multiple of the rms (estimated by the MADFM) above the mean (which for the wavelet arrays should be approximately zero). There are several other parameters that can be altered as well that affect the outcome of the reconstruction.

By default, the cube is reconstructed in three dimensions, using a 3-dimensional filter and 3-dimensional convolution. This can be altered, however, using the parameter `reconDim`. If set to 1, this means the cube is reconstructed by considering each spectrum separately, whereas `reconDim=2` will mean the cube is reconstructed by doing each channel map separately. The merits of these choices are discussed in §5, but it should be noted that a 2-dimensional reconstruction can be susceptible to edge effects if the spatial shape of the pixel array is not rectangular.

The user can also select the minimum scale to be used in the reconstruction. The first scale exhibits the highest frequency variations, and so ignoring this one can sometimes be beneficial in removing excess noise. The default is to use all scales (`minscale = 1`).

Finally, the filter that is used for the convolution can be selected by using `filterCode` and the relevant code number – the choices are listed in Appendix B. A larger filter will give a better reconstruction, but take longer and use more memory when executing. When multi-dimensional reconstruction is selected, this filter is used to construct a 2- or 3-dimensional equivalent.

3.4 Smoothing the cube

An alternative to doing the wavelet reconstruction is to smooth the cube. This technique can be useful in reducing the noise level slightly (at the cost of making neighbouring pixels correlated and blurring any signal present),

and is particularly well suited to the case where a particular signal size (i.e. a certain channel width or spatial size) is believed to be present in the data.

There are two alternative methods that can be used: spectral smoothing, using the Hanning filter; or spatial smoothing, using a 2D Gaussian kernel. These alternatives are outlined below. To utilise the smoothing option, set the parameter `flagSmooth=true` and set `smoothType` to either `spectral` or `spatial`.

3.4.1 Spectral smoothing

When `smoothType = spectral` is selected, the cube is smoothed only in the spectral domain. Each spectrum is independently smoothed by a Hanning filter, and then put back together to form the smoothed cube, which is then used by the searching algorithm (see below). Note that in the case of both the reconstruction and the smoothing options being requested, the reconstruction will take precedence and the smoothing will *not* be done.

There is only one parameter necessary to define the degree of smoothing – the Hanning width a (given by the user parameter `hanningWidth`). The coefficients $c(x)$ of the Hanning filter are defined by

$$c(x) = \begin{cases} \frac{1}{2}\left(1+\cos(\frac{\pi x}{a})\right) & |x| \leq (a+1)/2 \\ 0 & |x| > (a+1)/2. \end{cases}, \; a, x \in \mathbb{Z}$$

Note that the width specified must be an odd integer (if the parameter provided is even, it is incremented by one).

3.4.2 Spatial smoothing

When `smoothType = spatial` is selected, the cube is smoothed only in the spatial domain. Each channel map is independently smoothed by a two-dimensional Gaussian kernel, put back together to form the smoothed cube, and used in the searching algorithm (see below). Again, reconstruction is always done by preference if both techniques are requested.

The two-dimensional Gaussian has three parameters to define it, governed by the elliptical cross-sectional shape of the Gaussian function: the FWHM (full-width at half-maximum) of the major and minor axes, and the position angle of the major axis. These are given by the user parameters `kernMaj, kernMin` & `kernPA`. If a circular Gaussian is required, the user need only provide the `kernMaj` parameter. The `kernMin` parameter will then be set to the same value, and `kernPA` to zero. If we define these parameters as a, b, θ respectively, we can define the kernel by the function

$$k(x,y) = \exp\left[-0.5\left(\frac{X^2}{\sigma_X^2} + \frac{Y^2}{\sigma_Y^2}\right)\right]$$

where (x, y) are the offsets from the central pixel of the gaussian function, and

$$X = x\sin\theta - y\cos\theta \qquad Y = x\cos\theta + y\sin\theta$$
$$\sigma_X^2 = \frac{(a/2)^2}{2\ln 2} \qquad \sigma_Y^2 = \frac{(b/2)^2}{2\ln 2}$$

3.5 Input/Output of reconstructed/smoothed arrays

The smoothing and reconstruction stages can be relatively time-consuming, particularly for large cubes and reconstructions in 3-D (or even spatial smoothing). To get around this, *Duchamp* provides a shortcut to allow users to perform multiple searches (e.g. with different thresholds) on the same reconstruction/smoothing setup without re-doing the calculations each time.

To save the reconstructed array as a FITS file, set `flagOutputRecon = true`. The file will be saved in the same directory as the input image, so the user needs to have write permissions for that directory.

The filename will be derived from the input filename, with extra information detailing the reconstruction that has been done. For example, suppose `image.fits` has been reconstructed using a 3-dimensional reconstruction with filter #2, thresholded at 4σ using all scales. The output filename will then be `image.RECON-3-2-4-1.fits` (i.e. it uses the four parameters relevant for the *à trous* reconstruction as listed in Appendix B). The new FITS file will also have these parameters as header keywords. If a subsection of the input image has been used (see §3.1), the format of the output filename will be `image.sub.RECON-3-2-4-1.fits`, and the subsection that has been used is also stored in the FITS header.

Likewise, the residual image, defined as the difference between the input and reconstructed arrays, can also be saved in the same manner by setting `flagOutputResid = true`. Its filename will be the same as above, with `RESID` replacing `RECON`.

If a reconstructed image has been saved, it can be read in and used instead of redoing the reconstruction. To do so, the user should set the parameter `flagReconExists = true`. The user can indicate the name of the reconstructed FITS file using the `reconFile` parameter, or, if this is not specified, *Duchamp* searches for the file with the name as defined above. If the file is not found, the reconstruction is performed as normal. Note that to do this, the user needs to set `flagAtrous = true` (obviously, if this is `false`, the reconstruction is not needed).

To save the smoothed array, set `flagOutputSmooth = true`. The name of the saved file will depend on the method of smoothing used. It will be either `image.SMOOTH-1D-a.fits`, where a is replaced by the Hanning width used, or `image.SMOOTH-2D-a-b-c.fits`, where the Gaussian kernel

parameters are a,b,c. Similarly to the reconstruction case, a saved file can be read in by setting `flagSmoothExists = true` and either specifying a file to be read with the `smoothFile` parameter or relying on *Duchamp* to find the file with the name as given above.

3.6 Searching the image

3.6.1 Technique

The basic idea behind detection in *Duchamp* is to locate sets of contiguous voxels that lie above some threshold. No size or shape requirement is imposed upon the detections – that is, *Duchamp* does not fit e.g. a Gaussian profile to each source. All it does is find connected groups of bright voxels.

One threshold is calculated for the entire cube, enabling calculation of signal-to-noise ratios for each source (see Section 4 for details). The user can manually specify a value (using the parameter `threshold`) for the threshold, which will override the calculated value. Note that this only applies for the first of the two cases discussed below – the FDR case ignores any manually-set threshold value.

The cube is searched one channel map at a time, using the 2-dimensional raster-scanning algorithm of Lutz (1980) that connects groups of neighbouring pixels. Such an algorithm cannot be applied directly to a 3-dimensional case, as it requires that objects are completely nested in a row (when scanning along a row, if an object finishes and other starts, you won't get back to the first until the second is completely finished for the row). Three-dimensional data does not have this property, hence the need to treat the data on a 2-dimensional basis.

Although there are parameters that govern the minimum number of pixels in a spatial and spectral sense that an object must have (`minPix` and `minChannels` respectively), these criteria are not applied at this point. It is only after the merging and growing (see §3.7) is done that objects are rejected for not meeting these criteria.

Finally, the search only looks for positive features. If one is interested instead in negative features (such as absorption lines), set the parameter `flagNegative = true`. This will invert the cube (i.e. multiply all pixels by -1) prior to the search, and then re-invert the cube (and the fluxes of any detections) after searching is complete. All outputs are done in the same manner as normal, so that fluxes of detections will be negative.

3.6.2 Calculating statistics

A crucial part of the detection process is estimating the statistics that define the detection threshold. To determine a threshold, we need to estimate from the data two parameters: the middle of the noise distribution (the "noise

level"), and the width of the distribution (the "noise spread"). For both cases, we again use robust methods, using the median and MADFM.

The choice of pixels to be used depend on the analysis method. If the wavelet reconstruction has been done, the residuals (defined in the sense of original − reconstruction) are used to estimate the noise spread of the cube, since the reconstruction should pick out all significant structure. The noise level (the middle of the distribution) is taken from the original array.

If smoothing of the cube has been done instead, all noise parameters are measured from the smoothed array, and detections are made with these parameters. When the signal-to-noise level is quoted for each detection (see §4), the noise parameters of the original array are used, since the smoothing process correlates neighbouring pixels, reducing the noise level.

If neither reconstruction nor smoothing has been done, then the statistics are calculated from the original, input array.

The parameters that are estimated should be representative of the noise in the cube. For the case of small objects embedded in many noise pixels (e.g. the case of HI surveys), using the full cube will provide good estimators. It is possible, however, to use only a subsection of the cube by setting the parameter `flagStatSec = true` and providing the desired subsection to the `StatSec` parameter. This subsection works in exactly the same way as the pixel subsection discussed in §3.1. Note that this subsection applies only to the statistics used to determine the threshold. It does not affect the calculation of statistics in the case of the wavelet reconstruction.

3.6.3 Determining the threshold

Once the statistics have been calculated, the threshold is determined in one of two ways. The first way is a simple sigma-clipping, where a threshold is set at a fixed number n of standard deviations above the mean, and pixels above this threshold are flagged as detected. The value of n is set with the parameter `snrCut`. As before, the value of the standard deviation is estimated by the MADFM, and corrected by the ratio derived in Appendix G.

The second method uses the False Discovery Rate (FDR) technique (Hopkins et al. 2002; Miller et al. 2001), whose basis we briefly detail here. The false discovery rate (given by the number of false detections divided by the total number of detections) is fixed at a certain value α (e.g. $\alpha = 0.05$ implies 5% of detections are false positives). In practice, an α value is chosen, and the ensemble average FDR (i.e. $\langle FDR \rangle$) when the method is used will be less than α. One calculates p – the probability, assuming the null hypothesis is true, of obtaining a test statistic as extreme as the pixel value (the observed test statistic) – for each pixel, and sorts them in increasing order. One then calculates d where

$$d = \max_j \left\{ j : P_j < \frac{j\alpha}{c_N N} \right\},$$

and then rejects all hypotheses whose p-values are less than or equal to P_d. (So a $P_i < P_d$ will be rejected even if $P_i \geq j\alpha/c_N N$.) Note that "reject hypothesis" here means "accept the pixel as an object pixel" (i.e. we are rejecting the null hypothesis that the pixel belongs to the background).

The c_N value here is a normalisation constant that depends on the correlated nature of the pixel values. If all the pixels are uncorrelated, then $c_N = 1$. If N pixels are correlated, then their tests will be dependent on each other, and so $c_N = \sum_{i=1}^{N} i^{-1}$. Hopkins et al. (2002) consider real radio data, where the pixels are correlated over the beam. For the calculations done in *Duchamp*, $N = 2B$, where B is the beam size in pixels, calculated from the FITS header (if the correct keywords – BMAJ, BMIN – are not present, the size of the beam is taken from the parameter `beamSize`). The factor of 2 comes about because we treat neighbouring channels as correlated. In the case of a two-dimensional image, we just have $N = B$.

The theory behind the FDR method implies a direct connection between the choice of α and the fraction of detections that will be false positives. These detections, however, are individual pixels, which undergo a process of merging and rejection (§3.7), and so the fraction of the final list of detected objects that are false positives will be much smaller than α. See the discussion in §5.

3.7 Merging and growing detected objects

The searching step produces a list of detected objects that will have many repeated detections of a given object – for instance, spectral detections in adjacent pixels of the same object and/or spatial detections in neighbouring channels. These are then combined in an algorithm that matches all objects judged to be "close", according to one of two criteria.

One criterion is to define two thresholds – one spatial and one in velocity – and say that two objects should be merged if there is at least one pair of pixels that lie within these threshold distances of each other. These thresholds are specified by the parameters `threshSpatial` and `threshVelocity` (in units of pixels and channels respectively).

Alternatively, the spatial requirement can be changed to say that there must be a pair of pixels that are *adjacent* – a stricter, but perhaps more realistic requirement, particularly when the spatial pixels have a large angular size (as is the case for HI surveys). This method can be selected by setting the parameter `flagAdjacent` to 1 (i.e. `true`) in the parameter file. The velocity thresholding is done in the same way as the first option.

Once the detections have been merged, they may be "grown". This is a process of increasing the size of the detection by adding nearby pixels (according to the `threshSpatial` and `threshVelocity` parameters) that are above some secondary threshold. This threshold is lower than the one used for the initial detection, but above the noise level, so that faint pixels

are only detected when they are close to a bright pixel. The value of this threshold is a possible input parameter (`growthCut`), with a default value of 2σ.

The use of the growth algorithm is controlled by the `flagGrowth` parameter – the default value of which is `false`. If the detections are grown, they are sent through the merging algorithm a second time, to pick up any detections that now overlap or have grown over each other.

Finally, to be accepted, the detections must span *both* a minimum number of channels (enabling the removal of any spurious single-channel spikes that may be present), and a minimum number of spatial pixels. These numbers, as for the original detection step, are set with the `minChannels` and `minPix` parameters. The channel requirement means there must be at least one set of `minChannels` consecutive channels in the source for it to be accepted.

4 Outputs

4.1 During execution

Duchamp provides the user with feedback whilst it is running, to keep the user informed on the progress of the analysis. Most of this consists of self-explanatory messages about the particular stage the program is up to. The relevant parameters are printed to the screen at the start (once the file has been successfully read in), so the user is able to make a quick check that the setup is correct (see Appendix app-input for an example).

If the cube is being trimmed (§3.2), the resulting dimensions are printed to indicate how much has been trimmed. If a reconstruction is being done, a continually updating message shows either the current iteration and scale, compared to the maximum scale (when `reconDim = 3`), or a progress bar showing the amount of the cube that has been reconstructed (for smaller values of `reconDim`).

During the searching algorithms, the progress through the search is shown. When completed, the number of objects found is reported (this is the total number found, before any merging is done).

In the merging process (where multiple detections of the same object are combined – see §3.7), two stages of output occur. The first is when each object in the list is compared with all others. The output shows two numbers: the first being how far through the list the current object is, and the second being the length of the list. As the algorithm proceeds, the first number should increase and the second should decrease (as objects are combined). When the numbers meet, the whole list has been compared. If the objects are being grown, a similar output is shown, indicating the progress through the list. In the rejection stage, in which objects not meeting the minimum pixels/channels requirements are removed, the total number of objects remaining in the list is shown, which should steadily decrease with each rejection until all have been examined. Note that these steps can be very quick for small numbers of detections.

Since this continual printing to screen has some overhead of time and CPU involved, the user can elect to not print this information by setting the parameter `verbose = false`. In this case, the user is still informed as to the steps being undertaken, but the details of the progress are not shown.

There may also be Warning or Error messages printed to screen. The Warning messages occur when something happens that is unexpected (for instance, a desired keyword is not present in the FITS header), but not detrimental to the execution. An Error message is something more serious, and indicates some part of the program was not able to complete its task. The message will also indicate which function or subroutine generated it – this is primarily a tool for debugging, but can be useful in determining what went wrong.

4.2 Text-based output files

4.2.1 Table of results

Finally, we get to the results – the reason for running *Duchamp* in the first place. Once the detection list is finalised, it is sorted by the mean velocity of the detections (or, if there is no good WCS associated with the cube, by the mean z-pixel position). The results are then printed to the screen and to the output file, given by the `OutFile` parameter.

The output consists of three parts. First, a list of the parameters are printed to the output file, for future reference. Next, the detection threshold that was used is given, so comparison can be made with other searches. The noise level and its spread are also reported.

The most interesting part, however, is the list of detected objects. This list, an example of which can be seen in Appendix D, contains the following columns (note that the title of the columns depending on WCS information will depend on the details of the WCS projection: they are shown below for the Equatorial and Galactic projection cases).

`Obj#:`	The ID number of the detection (simply the sequential count for the list, which is ordered by increasing velocity, or channel number, if the WCS is not good enough to find the velocity).
`Name:`	The "IAU"-format name of the detection (derived from the WCS position – see below for a description of the format).
`X,Y,Z:`	The "centre" pixel position, determined by the input parameter `pixelCentre`.
`RA/GLON:`	The Right Ascension or Galactic Longitude of the centre of the object.
`DEC/GLAT:`	The Declination or Galactic Latitude of the centre of the object.
`VEL:`	The mean velocity of the object [units given by the `spectralUnits` parameter].
`w_RA/w_GLON:`	The width of the object in Right Ascension or Galactic Longitude (depending on FITS coordinates) [arcmin].
`w_DEC/w_GLAT:`	The width of the object in Declination Galactic Latitude [arcmin].
`w_VEL:`	The full velocity width of the detection (max channel − min channel, in velocity units [see note below]).
`F_int:`	The integrated flux over the object, in the units of flux times velocity, corrected for the beam if necessary.

`F_tot:` The sum of the flux values of all detected voxels.

`F_peak:` The peak flux over the object, in the units of flux.

`S/Nmax:` The signal-to-noise ratio at the peak pixel.

`X1, X2:` The minimum and maximum X-pixel coordinates.

`Y1, Y2:` The minimum and maximum Y-pixel coordinates.

`Z1, Z2:` The minimum and maximum Z-pixel coordinates.

`Npix:` The number of voxels (i.e. distinct (x, y, z) coordinates) in the detection.

`Flag:` Whether the detection has any warning flags (see below).

Note that the `X, Y, Z` columns depend on the `pixelCentre` parameter. This is because there are three alternative ways of expressing the centre of a detection, which are all listed in the list of detections written to the output file. These alternatives are:

`X_av, Y_av, Z_av:` The average pixel value in each axis direction i.e. X_av is the average of the x-values of all pixels in the detection.

`X_cent, Y_cent, Z_cent:` The centroid position, being the flux-weighted average of the pixels.

`X_peak, Y_peak, Z_peak:` The location of the pixel containing the peak flux value.

These are also written to the table in the output file, although not to the screen (as it would make the width of the table unwieldy). Similarly, the `F_tot` column is only written to the output file, and not at run-time.

The `Name` is derived from the WCS position. For instance, a source that is centred on the RA,Dec position $12^h53^m45^s$, -36°24′12″ will be given the name J125345−362412, if the epoch is J2000, or the name B125345−362412 if it is B1950. An alternative form is used for Galactic coordinates: a source centred on the position (l,b) = (323.1245, 5.4567) will be called G323.124+05.457. If the WCS is not valid (i.e. is not present or does not have all the necessary information), the `Name, RA, DEC, VEL` and related columns are not printed, but the pixel coordinates are still provided.

The velocity units can be specified by the user, using the parameter `spectralUnits` (enter it as a single string with no spaces). The default value is km/s, which should be suitable for most users. These units are also used to give the units of integrated flux. Note that if there is no rest frequency specified in the FITS header, the *Duchamp* output will instead default to using Frequency, with units of MHz.

If the WCS is absent or not sufficiently specified, then all columns from `RA/GLON` to `w_VEL` will be left blank. Also, `F_int` will be replaced with the more simple `F_tot`.

The `Flag` column contains any warning flags, such as:

- **E** – The detection is next to the spatial edge of the image, meaning either the limit of the pixels, or the limit of the non-BLANK pixel region.
- **S** – The detection lies at the edge of the spectral region.
- **N** – The total flux, summed over all the (non-BLANK) pixels in the smallest box that completely encloses the detection, is negative. Note that this sum is likely to include non-detected pixels. It is of use in pointing out detections that lie next to strongly negative pixels, such as might arise due to interference – the detected pixels might then also be due to the interference, so caution is advised.

4.2.2 Other results lists

Two additional results files can also be requested. One option is a VOTable-format XML file, containing just the RA, Dec, Velocity and the corresponding widths of the detections, as well as the fluxes. The user should set `flagVOT = true`, and put the desired filename in the parameter `votFile` – note that the default is for it not to be produced. This file should be compatible with all Virtual Observatory tools (such as Aladin[6] or TOPCAT[7]). The second option is an annotation file for use with the Karma toolkit of visualisation tools (in particular, with `kvis`). This will draw a circle at the position of each detection, scaled by the spatial size of the detection, and number it according to the Obj# given above. To make use of this option, the user should set `flagKarma = true`, and put the desired filename in the parameter `karmaFile` – again, the default is for it not to be produced.

As the program is running, it also (optionally) records the detections made in each individual spectrum or channel (see §3.6 for details on this process). This is recorded in the file given by the parameter `LogFile`. This file does not include the columns `Name, RA, DEC, w_RA, w_DEC, VEL, w_VEL`. This file is designed primarily for diagnostic purposes: e.g. to see if a given set of pixels is detected in, say, one channel image, but does not survive the merging process. The list of pixels (and their fluxes) in the final detection list are also printed to this file, again for diagnostic purposes. The file also records the execution time, as well as the command-line statement used to run *Duchamp*. The creation of this log file can be prevented by setting `flagLog = false` (which is the default).

[6] http://aladin.u-strasbg.fr/
[7] http://www.star.bristol.ac.uk/ mbt/topcat/

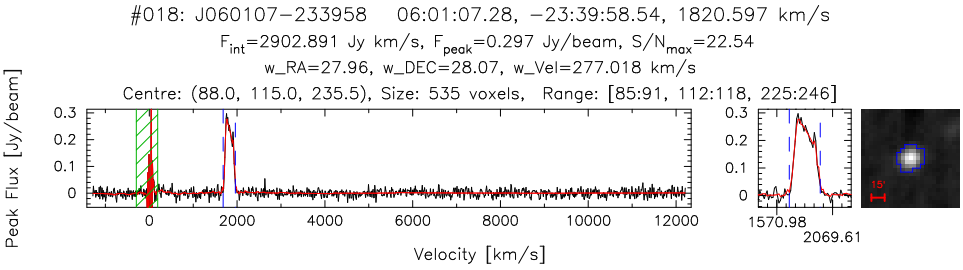


Figure 1: An example of the spectral output. Note several of the features discussed in the text: the red lines indicating the reconstructed spectrum; the blue dashed lines indicating the spectral extent of the detection; the green hashed area indicating the Milky Way channels that are ignored by the searching algorithm; the blue border showing its spatial extent on the 0th moment map; and the 15 arcmin-long scale bar.

4.3 Graphical output

4.3.1 Spectral plots

As well as the output data file, a postscript file is created that shows the spectrum for each detection, together with a small cutout image (the 0th moment) and basic information about the detection (note that any flags are printed after the name of the detection, in the format `[E]`). If the cube was reconstructed, the spectrum from the reconstruction is shown in red, over the top of the original spectrum. The spectral extent of the detected object is indicated by two dashed blue lines, and the region covered by the "Milky Way" channels is shown by a green hashed box. An example detection can be seen below in Fig. 1.

The spectrum that is plotted is governed by the `spectralMethod` parameter. It can be either `peak` (the default), where the spectrum is from the spatial pixel containing the detection's peak flux; or `sum`, where the spectrum is summed over all spatial pixels, and then corrected for the beam size. The spectral extent of the detection is indicated with blue lines, and a zoom is shown in a separate window.

The cutout image can optionally include a border around the spatial pixels that are in the detection (turned on and off by the `drawBorders` parameter – the default is `true`). It includes a scale bar in the bottom left corner to indicate size – its length is indicated next to it (the choice of length depends on the size of the image).

There may also be one or two extra lines on the image. A yellow line shows the limits of the cube's spatial region: when this is shown, the detected object will lie close to the edge, and the image box will extend outside the region covered by the data. A purple line, however, shows the dividing line between BLANK and non-BLANK pixels. The BLANK pixels will always be shown in black. The first type of line is always drawn, while the second

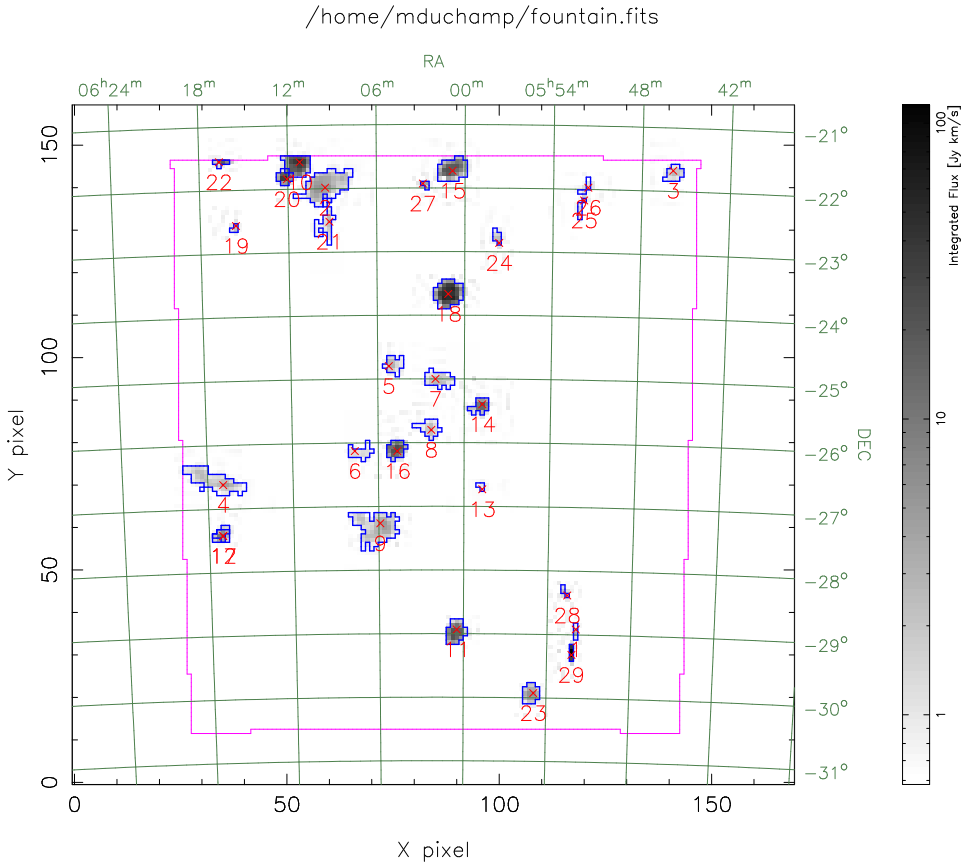


Figure 2: An example of the moment map created by *Duchamp*. The full extent of the cube is covered, and the 0th moment of each object is shown (integrated individually over all the detected channels). The purple line indicates the limit of the non-BLANK pixels.

is governed by the parameter `drawBlankEdges` (whose default is `true`), and obviously whether there are any BLANK pixel present.

4.3.2 Spatial maps

Finally, a couple of images are optionally produced: a 0th moment map of the cube, combining just the detected channels in each object, showing the integrated flux in grey-scale; and a "detection image", a grey-scale image where the pixel values are the number of channels that spatial pixel is detected in. In both cases, if `drawBorders = true`, a border is drawn around the spatial extent of each detection, and if `drawBlankEdges = true`, the purple line dividing BLANK and non-BLANK pixels (as described above) is drawn. An example moment map is shown in Fig. 2. The production or otherwise of these images is governed by the `flagMaps` parameter.

The moment map is also displayed in a PGPlot XWindow (with the `/xs` display option). This feature can be turned off by setting `flagXOutput = false` – this might be useful if running *Duchamp* on a terminal with no window display capability, or if you have set up a script to run it in a batch mode.

The purpose of these images are to provide a visual guide to where the detections have been made, and, particularly in the case of the moment map, to provide an indication of the strength of the source. In both cases, the detections are numbered (in the same sense as the output list and as the spectral plots), and the spatial borders are marked out as for the cutout images in the spectra file. Both these images are saved as postscript files (given by the parameters `momentMap` and `detectionMap` respectively), with the latter also displayed in a PGPLOT window (regardless of the state of `flagMaps`).

5 Notes and hints on the use of *Duchamp*

In using *Duchamp*, the user has to make a number of decisions about the way the program runs. This section is designed to give the user some idea about what to choose.

The main choice is whether to alter the cube to try and enhance the detectability of objects, by either smoothing or reconstructing via the *à trous* method. The main benefits of both methods are the marked reduction in the noise level, leading to regularly-shaped detections, and good reliability for faint sources.

The main drawback with the *à trous* method is the long execution time: to reconstruct a $170 \times 160 \times 1024$ (HIPASS) cube often requires three iterations and takes about 20-25 minutes to run completely. Note that this is for the more complete three-dimensional reconstruction: using `reconDim = 1` makes the reconstruction quicker (the full program then takes less than 5 minutes), but it is still the largest part of the time.

The smoothing procedure is computationally simpler, and thus quicker, than the reconstruction. The spectral Hanning method adds only a very small overhead on the execution, and the spatial Gaussian method, while taking longer, will be done (for the above example) in less than 2 minutes. Note that these times will depend on the size of the filter/kernel used: a larger filter means more calculations.

The searching part of the procedure is much quicker: searching an unreconstructed cube leads to execution times of less than a minute. Alternatively, using the ability to read in previously-saved reconstructed arrays makes running the reconstruction more than once a more feasible prospect.

On the positive side, the shape of the detections in a cube that has been reconstructed or smoothed will be much more regular and smooth – the ragged edges that objects in the raw cube possess are smoothed by the removal of most of the noise. This enables better determination of the shapes and characteristics of objects.

While the time overhead is larger for the reconstruction case, it will potentially provide a better recovery of real sources than the smoothing case. This is because it probes the full range of scales present in the cube (or spectral domain), rather than the specific scale determined by the Hanning filter or Gaussian kernel used in the smoothing.

When considering the reconstruction method, note that the 2D reconstruction (`reconDim = 2`) can be susceptible to edge effects. If the valid area in the cube (i.e. the part that is not BLANK) has non-rectangular edges, the convolution can produce artefacts in the reconstruction that mimic the edges and can lead (depending on the selection threshold) to some spurious sources. Caution is advised with such data – the user is advised to check carefully the reconstructed cube for the presence of such artefacts. Note, however, that the 1- and 3-dimensional reconstructions are *not* susceptible

in the same way, since the spectral direction does not generally exhibit these BLANK edges, and so we recommend the use of either of these.

If one chooses the reconstruction method, a further decision is required on the signal-to-noise cutoff used in determining acceptable wavelet coefficients. A larger value will remove more noise from the cube, at the expense of losing fainter sources, while a smaller value will include more noise, which may produce spurious detections, but will be more sensitive to faint sources. Values of less than about 3σ tend to not reduce the noise a great deal and can lead to many spurious sources (this depends, of course on the cube itself).

The smoothing options have less parameters to consider: basically just the size of the smoothing function or kernel. Spectrally smoothing with a Hanning filter of width 3 (the smallest possible) is very efficient at removing spurious one-channel objects that may result just from statistical fluctuations of the noise. One may want to use larger widths or kernels of larger size to look for features of a particular scale in the cube.

When it comes to searching, the FDR method produces more reliable results than simple sigma-clipping, particularly in the absence of reconstruction. However, it does not work in exactly the way one would expect for a given value of `alpha`. For instance, setting fairly liberal values of `alpha` (say, 0.1) will often lead to a much smaller fraction of false detections (i.e. much less than 10%). This is the effect of the merging algorithms, that combine the sources after the detection stage, and reject detections not meeting the minimum pixel or channel requirements. It is thus better to aim for larger `alpha` values than those derived from a straight conversion of the desired false detection rate.

If the FDR method is not used, caution is required when choosing the S/N cutoff. Typical cubes have very large numbers of pixels, so even an apparently large cutoff will still result in a not-insignificant number of detections simply due to random fluctuations of the noise background. For instance, a 4σ threshold on a cube of Gaussian noise of size $100 \times 100 \times 1024$ will result in ~ 340 detections. This is where the minimum channel and pixel requirements are important in rejecting spurious detections.

Finally, as *Duchamp* is still undergoing development, there are some elements that are not fully developed. In particular, it is not as clever as I would like at avoiding interference. The ability to place requirements on the minimum number of channels and pixels partially circumvents this problem, but work is being done to make *Duchamp* smarter at rejecting signals that are clearly (to a human eye at least) interference. See the following section for further improvements that are planned.

6 Future developments

Here are lists of planned improvements and a wish-list of features that would be nice to include (but are not planned in the immediate future). Let me know if there are items not on these lists, or items on the list you would like prioritised.

Planned developments:

- Parallelisation of the code, to improve speed particularly on multi-core machines.
- Better determination of the noise characteristics of spectral-line cubes, including understanding how the noise is generated and developing a model for it.
- Include more source analysis. Examples could be: shape information; measurements of HI mass; more variety of measurements of velocity width and profile.
- Improved ability to reject interference, possibly on the spectral shape of features.
- Ability to separate (de-blend) distinct sources that have been merged.

Wish-list:

- Incorporation of Swinburne's S2PLOT [8] code for improved visualisation.
- Link to lists of possible counterparts (e.g. via NED/SIMBAD/other VO tools?).
- On-line web service interface, so a user can upload a cube and get back a source-list.
- Embed *Duchamp* in a GUI, to move away from the text-based interaction.

7 Why "*Duchamp*"?

Well, it's important for a program to have a name, and the initial working title of *cubefind* was somewhat uninspiring. I wanted to avoid the classic astronomical approach of designing a cute acronym, and since it is designed to work on cubes, I looked at naming it after a cubist. *Picasso*, sadly, was already taken (Minchin 1999), so I settled on naming it after Marcel Duchamp, another cubist, but also one of the first artists to work with "found objects".

[8] http://astronomy.swin.edu.au/s2plot/

References

Calabretta M., Greisen E., 2002, A&A, 395, 1077

Greisen E., Calabretta M., 2002, A&A, 395, 1061

Hanisch R., Farris A., Greisen E., Pence W., Schlesinger B., Teuben P., Thompson R., Warnock A., 2001, A&A, 376, 359

Hopkins A., Miller C., Connolly A., Genovese C., Nichol R., Wasserman L., 2002, AJ, 123, 1086

Lutz R., 1980, The Computer Journal, 23, 262

Meyer M., et al., 2004, MNRAS, 350, 1195

Miller C., Genovese C., Nichol R., Wasserman L., Connolly A., Reichart D., Hopkins A., Schneider J., Moore A., 2001, AJ, 122, 3492

Minchin R., 1999, PASA, 16, 12

Starck J.-L., Murtagh F., 2002, "Astronomical Image and Data Analysis". Springer

A Obtaining and installing *Duchamp*

A.1 Installing

The *Duchamp* web page can be found at the following location:
http://www.atnf.csiro.au/people/Matthew.Whiting/Duchamp
Here you can find a gzipped tar archive of the source code that can be downloaded and extracted, as well as this User's Guide in postscript and hyperlinked PDF formats.

To build *Duchamp*, you will need three main external libraries: PGPLOT, CFITSIO (this needs to be version 2.5 or greater – version 3+ is better) and WCSLIB. If these are not present on your system, you can download them from the following locations:

- PGPLOT: http://www.astro.caltech.edu/ tjp/pgplot/
- CFITSIO: http://heasarc.gsfc.nasa.gov/docs/software/fitsio/fitsio.html
- WCSLIB: http://www.atnf.csiro.au/people/Mark.Calabretta/WCS/index.html

Duchamp can be built on Unix/Linux systems by typing (assuming that the prompt your terminal provides is a `>` – don't type this character!):

```
> ./configure
> make
> make clean (optional -- to remove the object files)
```

Run in this manner, `configure` should find all the necessary libraries, but if some libraries have been installed in non-standard locations, it may fail. In this case, you can specify additional directories to look in by giving extra command-line arguments. There are separate options for library files (eg. libcpgplot.a) and header files (eg. cpgplot.h).

For example, suppose WCSLIB had been locally installed in the directory `/home/mduchamp/wcslib`. There will then be two libraries created that are likely to be in the following subdirectories: `C/` and `pgsbox/`. Each subdirectory needs to be searched for library and header files, so one could build Duchamp by typing:

```
> ./configure \
LIBDIRS="/home/mduchamp/wcslib/C /home/mduchamp/wcslib/pgsbox" \
INCDIRS="/home/mduchamp/wcslib/C /home/mduchamp/wcslib/pgsbox"
```

And then just run make in the usual fashion:

```
> make
```

This will produce the executable `Duchamp`. You can verify that it is running correctly by running the verification shell script:

```
> VerifyDuchamp.sh
```

This will use a dummy FITS image in the `verification/` directory – this image has some Gaussian random noise, with five Gaussian sources present, plus a dummy WCS. The script runs Duchamp on this image with four different sets of inputs, and compares to known results, looking for differences and reporting any. There should be none reported if everything is working correctly.

A.2 Running *Duchamp*

You can then run *Duchamp* on your own data. This can be done in one of two ways. The first is:

```
> Duchamp -f [FITS file]
```

where `[FITS file]` is the file you wish to search. This method simply uses the default values of all parameters.

The second method allows some determination of the parameter values by the user. Type:

```
> Duchamp -p [parameter file]
```

where `[parameterFile]` is a file with the input parameters, including the name of the cube you want to search. There are two example input files included with the distribution. The smaller one, `InputExample`, shows the typical parameters one might want to set. The large one, `InputComplete`, lists all possible parameters that can be entered, and a brief description of them. To get going quickly, just replace the `"your-file-here"` in the `InputExample` file with your image name, and type

```
> Duchamp -p InputExample
```

To disable the use of X-window plotting (in displaying the map of detections), one can either set the parameter `flagXOutput = false` or use the `-x` command-line option:

```
> Duchamp -x -p [parameter file] , or
> Duchamp -x -f [FITS file]
```

The following appendices provide details on the individual parameters, and show examples of the output files that *Duchamp* produces.

A.3 Feedback

It may happen that you discover bugs or problems with *Duchamp*, or you have suggestions for improvements or additional features to be included in future releases. You can submit a "ticket" (a trackable bug report) at the

Duchamp Trac wiki at the following location:
http://sourcecode.atnf.csiro.au/cgi-bin/trac_duchamp.cgi/newticket
(there is a link to this page from the Duchamp website).

There is also an email exploder, duchamp-user[**at**]atnf.csiro.au, that users can subscribe to keep up to date with changes, updates, and other news about *Duchamp*. To subscribe, send an email (from the account you wish to subscribe to the list) to duchamp-user-request[**at**]atnf.csiro.au with the single word "subscribe" in the body of the message. To be removed from this list, send a message with "unsubscribe" in its body to the same address.

A.4 Beta Versions

On the *Duchamp* website there may be a beta version listed in the downloads section. As *Duchamp* is still under development, there will be times when there has been new functionality added to the code, but the time has not yet come to release a new minor (or indeed major) version.

Sometimes I will post the updated version of the code on the website as a "beta" version, particularly if I'm interested in people testing it. It will not have been tested as rigorously as the proper releases, but it will certainly work in the basic cases that I use to test it during development. So feel free to give it a try – the `CHANGES` file will usually detail what is different to the last numbered release.

B Available parameters

The full list of parameters that can be listed in the input file are given here. If not listed, they take the default value given in parentheses. Since the order of the parameters in the input file does not matter, they are grouped here in logical sections.

Input related

`ImageFile [no default]`:
: The filename of the data cube to be analysed.

`flagSubsection [false]`:
: A flag to indicate whether one wants a subsection of the requested image.

`Subsection [ [*,*,*] ]`:
: The requested subsection – see §3.1 for details on the subsection format. If the full range of a dimension is required, use a `*` (thus the default is the full cube).

`flagReconExists [false]`:
: A flag to indicate whether the reconstructed array has been saved by a previous run of *Duchamp*. If set true, the reconstructed array will be read from the file given by `reconFile`, rather than calculated directly.

`reconFile [no default]`:
: The FITS file that contains the reconstructed array. If `flagReconExists` is true and this parameter is not defined, the default file that is looked for will be determined by the *à trous* parameters (see §3.3).

`flagSmoothExists [false]`:
: A flag to indicate whether the Hanning-smoothed array has been saved by a previous run of *Duchamp*. If set true, the smoothed array will be read from the file given by `smoothFile`, rather than calculated directly.

`smoothFile [no default]`:
: The FITS file that has a previously smoothed array. If `flagSmoothExists` is true and this parameter is not defined, the default file that is looked for will be determined by the smoothing parameters (see §3.4).

Output related

`OutFile [duchamp-Results.txt]`:
: The file containing the final list of detections. This also records the list of input parameters.

`SpectraFile [duchamp-Spectra.ps]`:
The postscript file containing the resulting integrated spectra and images of the detections.

`flagLog [false]`: A flag to indicate whether the details of intermediate detections should be logged.

`LogFile [duchamp-Logfile.txt]`:
The file in which intermediate detections are logged. These are detections that have not been merged. This is primarily for use in debugging and diagnostic purposes: normal use of the program will probably not require it.

`flagOutputRecon [false]`:
A flag to say whether or not to save the reconstructed cube as a FITS file. The filename will be derived according to the naming scheme detailed in Section 3.5.

`flagOutputResid [false]`:
As for `flagOutputRecon`, but for the residual array – the difference between the original cube and the reconstructed cube. The filename will be derived according to the naming scheme detailed in Section 3.5.

`flagOutputSmooth [false]`:
A flag to say whether or not to save the smoothed cube as a FITS file. The filename will be derived according to the naming scheme detailed in Section 3.4.

`flagVOT [false]`: A flag to say whether to create a VOTable file with the detection information. This will be an XML file in the Virtual Observatory VOTable format.

`votFile [duchamp-Results.xml]`:
The VOTable file with the list of final detections. Some input parameters are also recorded.

`flagKarma [false]`:
A flag to say whether to create a Karma annotation file corresponding to the information in `outfile`. This can be used as an overlay in Karma programs such as `kvis`.

`karmaFile [duchamp-Results.ann]`:
The Karma annotation file showing the list of final detections.

`flagMaps [true]`: A flag to say whether to save postscript files showing the 0th moment map of the whole cube (parameter `momentMap`) and the detection image (`detectionMap`).

`momentMap [duchamp-MomentMap.ps]`:
A postscript file containing a map of the 0th moment of

the detected sources, as well as pixel and WCS coordinates.

`detectionMap [duchamp-DetectionMap.ps]`:
A postscript file with a map showing each of the detected objects, coloured in greyscale by the number of detected channels in each spatial pixel. Also shows pixel and WCS coordinates.

`flagXOutput [true]`:
A flag to say whether to display a 0th moment map in a PGPlot X-window. This will be in addition to any that are saved to a file. This parameter can be overridden by the use of the `-x` command-line option, which disables the X-windows output.

Modifying the cube

`flagTrim [false]`:
A flag to say whether to trim BLANK pixels from the edges of the cube – these are typically pixels set to some particular value because they fall outside the imaged area, and trimming them can help speed up the execution.

`flagMW [false]`: A flag to say whether to ignore channels contaminated by Milky Way (or other) emission – the searching algorithms will not look at these channels.

`maxMW [112]`: The maximum channel number that contains "Milky Way" emission.

`minMW [75]`: The minimum channel number that contains "Milky Way" emission. Note that the range specified by `maxMW` and `minMW` is inclusive.

`flagBaseline [false]`:
A flag to say whether to remove the baseline from each spectrum in the cube for the purposes of reconstruction and detection.

Detection related

General detection

`flagStatSec [false]`:
A flag indicating whether the statistics should be calculated on a subsection of the cube, rather than the full cube. Note that this only applies to the statistics used to determine the threshold, and not for other statistical

calculations (such as those in the reconstruction phase).

`StatSec [ [*,*,*] ]`:
The subsection of the cube used for calculating statistics – see §3.1 for details on the subsection format. Only used if `flagStatSec=true`.

`flagNegative [false]`:
A flag indicating that the features of interest are negative. The cube is inverted prior to searching.

`snrCut [3.]`: The threshold, in multiples of σ above the mean.

`threshold [no default]`:
The actual value of the threshold. Normally the threshold is calculated from the cube's statistics, but the user can manually specify a value to be used that overrides the calculated value. If this is not specified, the calculated value is used. Also, when the FDR method is requested (see below), the value of the `threshold` parameter is ignored.

`flagGrowth [false]`:
A flag indicating whether or not to grow the detected objects to a smaller threshold.

`growthCut [2.]`: The smaller threshold using in growing detections. In units of σ above the mean.

`beamSize [10.]`: The size of the beam in pixels. If the header keywords BMAJ and BMIN are present, then these will be used to calculate the beam size, and this parameter will be ignored.

À trous reconstruction

`flagATrous [false]`:
A flag indicating whether or not to reconstruct the cube using the *à trous* wavelet reconstruction. See §3.3 for details.

`reconDim [1]`: The number of dimensions to use in the reconstruction. 1 means reconstruct each spectrum separately, 2 means each channel map is done separately, and 3 means do the whole cube in one go.

`scaleMin [1]`: The minimum wavelet scale to be used in the reconstruction. A value of 1 means "use all scales".

`snrRecon [4]`: The thresholding cutoff used in the reconstruction – only wavelet coefficients this many σ above the mean (or

greater) are included in the reconstruction.

`filterCode [1]:` The code number of the filter to use in the reconstruction. The options are:

- **•1:** B_3-spline filter: coefficients = $(\frac{1}{16}, \frac{1}{4}, \frac{3}{8}, \frac{1}{4}, \frac{1}{16})$
- **•2:** Triangle filter: coefficients = $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$
- **•3:** Haar wavelet: coefficients = $(0, \frac{1}{2}, \frac{1}{2})$

Smoothing the cube

`flagSmooth [false]:`
A flag indicating whether to smooth the cube. See §3.4 for details.

`smoothType [spectral]:`
The smoothing method used: either "spectral" (with a 1D Hanning filter) or "spatial" (with a 2D Gaussian filter).

`hanningWidth [5]:`
The width of the Hanning smoothing kernel.

`kernMaj [3]:` The full-width-half-maximum (FWHM) of the 2D Gaussian smoothing kernel's major axis.

`kernMin [3]:` The FWHM of the 2D Gaussian smoothing kernel's minor axis.

`kernPA [0]:` The position angle, in degrees, anticlockwise from vertical (i.e. usually East of North).

FDR method

`flagFDR [false]:` A flag indicating whether or not to use the False Discovery Rate method in thresholding the pixels.

`alphaFDR [0.01]:` The α parameter used in the FDR analysis. The average number of false detections, as a fraction of the total number, will be less than α (see §3.6).

Merging detections

`minPix [2]:` The minimum number of spatial pixels for a single detection to be counted.

`minChannels [3]:` At least one contiguous set of this many channels must be present in the detection for it to be accepted.

`flagAdjacent [true]:`
A flag indicating whether to use the "adjacent pixel" criterion to decide whether to merge objects. If not,

the next two parameters are used to determine whether objects are within the necessary thresholds.

`threshSpatial [3.]`:
The maximum allowed minimum spatial separation (in pixels) between two detections for them to be merged into one. Only used if `flagAdjacent = false`.

`threshVelocity [7.]`:
The maximum allowed minimum channel separation between two detections for them to be merged into one.

Other parameters

`spectralMethod [peak]`:
This indicates which method is used to plot the output spectra: `peak` means plot the spectrum containing the detection's peak pixel; `sum` means sum the spectra of each detected spatial pixel, and correct for the beam size. Any other choice defaults to `peak`.

`spectralUnits [km/s]`:
The user can specify the units of the spectral axis. Assuming the WCS of the FITS file is valid, the spectral axis is transformed into velocity, and put into these units for all output and for calculations such as the integrated flux of a detection.

`pixelCentre [centroid]`:
Which of the three ways of expressing the "centre" of a detection (see §4.2.1 for a description of the options) to use for the X, Y, & Z columns in the output list. Alternatives are: `centroid, peak, average`.

`drawBorders [true]`:
A flag indicating whether to draw borders around the detected objects in the moment maps included in the output (see for example Fig. 1).

`drawBlankEdges [true]`:
A flag indicating whether to draw the dividing line between BLANK and non-BLANK pixels on the 2D images (see for example Fig. 2).

`verbose [true]`: A flag indicating whether to print the progress of any computationally intensive algorithms (e.g. reconstruction, searching or merging algorithms) to the screen.

C Example parameter files

This is what a typical parameter file would look like.

```
imageFile       /home/mduchamp/fountain.fits
logFile         logfile.txt
outFile         results.txt
spectraFile     spectra.ps
flagSubsection  false
flagOutputRecon false
flagOutputResid 0
flagTrim        1
flagMW          1
minMW           75
maxMW           112
flagGrowth      1
growthCut       1.5
flagATrous      1
reconDim        1
scaleMin        1
snrRecon        4
flagFDR         1
alphaFDR        0.1
snrCut          3
threshSpatial   3
threshVelocity  7
```

Note that, as in this example, the flag parameters can be entered as strings (`true`/`false`) or integers (`1`/`0`). Also, note that it is not necessary to include all these parameters in the file, only those that need to be changed from the defaults (as listed in Appendix B), which in this case would be very few. A minimal parameter file might look like:

```
imageFile       /home/mduchamp/fountain.fits
flagLog         false
flagATrous      1
snrRecon        3
snrCut          2.5
minChannels     4
```

This will reconstruct the cube with a lower SNR value than the default, select objects at a lower threshold, with a looser minimum channel requirement, and not keep a log of the intermediate detections.

The following page demonstrates how the parameters are presented to the user, both on the screen at execution time, and in the output and log

files. On each line, there is a description on the parameter, the relevant parameter name that is used in the input file (if there is one that the user can enter), and the value of the parameter being used.

```
---- Parameters ----
Image to be analysed.....................[imageFile]  =  fountain.fits
Intermediate Logfile.....................[logFile]  =  duchamp-Logfile.txt
Final Results file.......................[outFile]  =  duchamp-Results.txt
Spectrum file............................[spectraFile]  =  duchamp-Spectra.ps
Oth Moment Map...........................[momentMap]  =  duchamp-MomentMap.ps
Detection Map............................[detectionMap]  =  duchamp-DetectionMap.ps
Display a map in a pgplot xwindow?.......[flagXOutput]  =  true
Saving reconstructed cube?...............[flagoutputrecon]  =  false
Saving residuals from reconstruction?..[flagoutputresid]  =  false
------
Blank Pixel Value.....................................  =  -8.00061
Trimming Blank Pixels?...................[flagTrim]  =  true
Searching for Negative features?.........[flagNegative]  =  false
Removing Milky Way channels?.............[flagMW]  =  true
Milky Way Channels.....................[minMW - maxMW]  =  75-112
Beam Size (pixels)....................................  =  10.1788
Removing baselines before search?........[flagBaseline]  =  false
Smoothing each spectrum first?...........[flagSmooth]  =  false
Using A Trous reconstruction?............[flagATrous]  =  true
Number of dimensions in reconstruction.....[reconDim]  =  1
Minimum scale in reconstruction............[scaleMin]  =  1
SNR Threshold within reconstruction........[snrRecon]  =  4
Filter being used for reconstruction.....[filterCode]  =  1 (B3 spline function)
Using FDR analysis?......................[flagFDR]  =  false
SNR Threshold (in sigma).....................[snrCut]  =  3
Minimum # Pixels in a detection..............[minPix]  =  2
Minimum # Channels in a detection.........[minChannels]  =  3
Growing objects after detection?..........[flagGrowth]  =  false
Using Adjacent-pixel criterion?..........[flagAdjacent]  =  true
Max. velocity separation for merging....[threshVelocity]  =  7
Method of spectral plotting.............[spectralMethod]  =  peak
Type of object centre used in results......[pixelCentre]  =  centroid
--------------------
```

D Example results file

This the typical content of an output file, after running *Duchamp* with the parameters illustrated on the previous page. The table is split over two pages so that it can fit.

```
Results of the Duchamp source finder: Thu May  3 12:15:44 2007

---- Parameters ----
      (... omitted for clarity -- see previous page for examples...)
--------------------
Summary of statistics:
  Detection threshold = 0.0396039 Jy/beam
  Noise level = 0.000122074, Noise spread = 0.0131606
Total number of detections = 29
--------------------
```

Obj#	Name	X	Y	Z	RA	DEC	VEL [km/s]	w_RA [arcmin]	w_DEC [arcmin]	w_VEL [km/s]	F_int [Jy km/s]	F_tot [Jy/beam]	F_peak [Jy/beam]
1	J055156-285617	118.0	35.6	73.1	05:51:56.48	-28:56:17.66	-322.567	3.66	16.10	26.383	2.769	0.534	0.123
2	J060928-215732	59.0	140.4	114.5	06:09:28.88	-21:57:32.05	224.545	56.48	35.27	65.957	1080.700	12.635	0.213
3	J054549-214307	141.3	143.3	114.7	05:45:49.83	-21:43:07.02	227.076	19.61	16.66	26.383	29.090	1.603	0.090
4	J061734-263243	32.8	71.0	115.5	06:17:34.22	-26:32:43.65	237.116	60.93	26.21	26.383	442.684	6.445	0.117
5	J060457-244631	74.9	98.3	116.3	06:04:57.69	-24:46:31.96	247.970	20.11	19.89	39.574	41.062	1.980	0.109
6	J060717-260846	67.2	77.7	117.0	06:07:17.01	-26:08:46.62	257.675	24.20	19.76	26.383	25.700	1.322	0.094
7	J060144-250045	85.9	94.8	117.9	06:01:44.91	-25:00:45.16	268.817	27.99	20.02	26.383	78.836	2.645	0.124
8	J060224-254720	83.6	83.2	118.0	06:02:24.45	-25:47:20.55	269.695	28.03	19.97	26.383	42.285	1.813	0.118
9	J060607-271947	71.2	60.0	121.3	06:06:07.79	-27:19:47.64	314.376	48.35	35.61	26.383	748.544	9.025	0.150
10	J061118-213640	52.5	145.5	162.5	06:11:18.67	-21:36:40.65	857.911	28.33	19.55	118.722	1196.068	31.825	0.410
11	J060034-285856	89.7	35.3	202.3	06:00:34.42	-28:58:56.16	1382.442	19.92	24.09	197.870	480.427	16.118	0.173
12	J061702-272242	35.0	58.6	216.8	06:17:02.06	-27:22:42.42	1573.375	16.51	15.53	52.765	16.193	1.388	0.078
13	J055848-264047	95.7	69.8	223.3	05:58:48.74	-26:40:47.93	1658.867	7.95	8.05	39.574	1.035	0.266	0.062
14	J055848-252526	95.8	88.6	232.2	05:58:48.82	-25:25:26.30	1776.793	19.91	16.11	237.444	118.175	7.014	0.115
15	J060053-214225	88.9	144.3	233.0	06:00:53.22	-21:42:25.46	1787.741	27.95	24.13	211.061	924.853	20.989	0.166
16	J060444-260635	75.8	78.3	233.3	06:04:44.38	-26:06:35.38	1790.634	20.11	19.90	237.444	334.362	15.177	0.155
17	J061707-272459	34.7	58.0	234.8	06:17:07.93	-27:24:59.96	1811.015	16.39	11.53	105.531	37.022	2.597	0.093
18	J060107-233958	88.0	115.0	235.5	06:01:07.28	-23:39:58.54	1820.597	27.96	28.07	277.018	2902.891	58.946	0.297
19	J061536-223539	37.9	130.4	254.4	06:15:36.86	-22:35:39.07	2069.235	8.19	7.80	52.765	2.317	0.596	0.069
20	J061209-214914	49.6	142.3	269.9	06:12:09.67	-21:49:14.61	2273.533	16.27	15.73	382.548	151.961	9.771	0.101
21	J060925-223205	59.3	131.8	295.5	06:09:25.43	-22:32:05.29	2611.227	20.60	47.80	26.383	65.724	2.113	0.177
22	J061622-213304	34.9	146.0	298.5	06:16:22.96	-21:33:04.02	2650.714	16.21	7.58	158.296	19.000	2.932	0.127
23	J055505-295615	107.4	20.8	367.5	05:55:05.81	-29:56:15.45	3561.657	15.70	20.26	52.765	84.750	3.847	0.169
24	J055743-224809	99.8	127.9	433.1	05:57:43.20	-22:48:09.38	4427.169	7.89	16.09	197.870	10.467	1.346	0.167
25	J055204-221558	119.4	135.6	433.9	05:52:04.65	-22:15:58.54	4437.091	7.68	20.19	52.765	4.725	0.729	0.173
26	J055147-215849	120.5	139.9	434.2	05:51:47.19	-21:58:49.72	4441.344	11.74	16.27	26.382	4.701	0.725	0.181
27	J060246-215620	82.3	140.9	434.5	06:02:46.59	-21:56:20.03	4445.066	8.01	8.01	197.870	5.576	1.434	0.294
28	J055245-282025	115.5	44.6	647.7	05:52:45.55	-28:20:25.30	7257.209	7.77	12.15	65.957	2.999	0.771	0.263
29	J055212-291701	117.0	30.5	727.0	05:52:12.81	-29:17:01.01	8303.959	3.67	16.10	356.166	143.069	27.599	0.479

... continued ...

S/Nmax	X1	X2	Y1	Y2	Z1	Z2	Npix [pix]	Flag	X_av	Y_av	Z_av	X_cent	Y_cent	Z_cent	X_peak	Y_peak	Z_peak
9.37	118	118	34	37	72	74	6		118.0	35.7	73.0	118.0	35.6	73.1	118	36	73
16.14	52	65	136	144	113	118	151		59.1	140.4	114.6	59.0	140.4	114.5	59	140	114
6.82	139	143	142	145	114	116	24		141.3	143.3	114.7	141.3	143.3	114.7	141	144	115
8.90	26	40	68	74	115	117	81	E	32.8	71.1	115.5	32.8	71.0	115.5	35	70	115
8.26	73	77	96	100	115	118	27		74.9	98.4	116.4	74.9	98.3	116.3	74	98	116
7.14	65	70	76	80	116	118	20		67.2	77.7	117.0	67.2	77.7	117.0	66	78	117
9.41	83	89	93	97	117	119	36		85.9	94.8	117.9	85.9	94.8	117.9	85	95	118
8.95	80	86	81	85	117	119	24		83.6	83.1	117.9	83.6	83.2	118.0	84	83	118
11.36	65	76	55	63	120	122	115		71.2	60.0	121.3	71.2	60.0	121.3	72	61	121
31.17	49	55	143	147	158	167	225	E	52.5	145.3	162.5	52.5	145.5	162.5	53	146	164
13.13	88	92	33	38	195	210	218		89.7	35.3	202.4	89.7	35.3	202.3	90	36	197
5.94	33	36	57	60	215	219	23		35.0	58.6	216.9	35.0	58.6	216.8	35	58	215
4.68	95	96	69	70	222	225	6		95.7	69.8	223.3	95.7	69.8	223.3	96	69	223
8.77	93	97	87	90	221	239	115		95.7	88.6	231.8	95.8	88.6	232.2	96	89	237
12.57	86	92	142	147	223	239	280	E	88.8	144.3	232.5	88.9	144.3	233.0	89	144	233
11.73	74	78	76	80	224	242	218		75.8	78.4	233.2	75.8	78.3	233.3	76	78	240
7.05	33	36	57	59	229	237	44		34.7	58.1	234.5	34.7	58.0	234.8	35	58	236
22.54	85	91	112	118	225	246	535		88.0	115.0	235.8	88.0	115.0	235.5	88	115	231
5.24	37	38	130	131	252	256	11		37.9	130.5	254.2	37.9	130.4	254.4	38	131	256
7.65	48	51	141	144	257	286	175		49.6	142.3	270.3	49.6	142.3	269.9	50	142	259
13.45	57	61	127	138	295	297	25		59.1	132.0	295.6	59.3	131.8	295.5	60	132	295
9.65	33	36	145	146	293	305	40	E	34.9	146.0	298.8	34.9	146.0	298.5	34	146	297
12.85	106	109	19	23	366	370	42		107.5	20.9	367.6	107.4	20.8	367.5	108	21	367
12.71	99	100	127	130	423	438	17	N	99.7	128.1	433.3	99.8	127.9	433.1	100	127	432
13.17	119	120	133	137	431	435	8		119.5	135.8	433.6	119.4	135.6	433.9	120	137	434
13.77	119	121	139	142	433	435	8		120.5	140.0	434.2	120.5	139.9	434.2	121	140	434
22.35	82	83	140	141	425	440	13		82.5	140.8	434.2	82.3	140.9	434.5	82	141	436
19.99	115	116	44	46	645	650	8		115.4	44.8	647.4	115.5	44.6	647.7	116	44	648
36.36	117	117	29	32	714	741	104		117.0	30.5	727.3	117.0	30.5	727.0	117	30	733

Note that the width of the table can make it hard to read. A good trick for those using UNIX/Linux is to make use of the `a2ps` command. The following works well, producing a postscript file `duchamp-Results.ps`:

```
a2ps -1 -r -f5 -o duchamp-Results.ps duchamp-Results.txt
```

E Example VOTable output

This is part of the VOTable, in XML format, corresponding to a basic FDR search.

```
<?xml version="1.0"?>
<VOTABLE version="1.1" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:noNamespaceSchemaLocation="http://www.ivoa.net/xml/VOTable/VOTable/v1.1">
  <COOSYS ID="J2000" equinox="J2000." epoch="J2000." system="eq_FK5"/>
  <RESOURCE name="Duchamp Output">
    <TABLE name="Detections">
      <DESCRIPTION>Detected sources and parameters from running the Duchamp source finder.</DESCRIPTION>
      <PARAM name="FITS file" datatype="char" ucd="meta.file;meta.fits" value="/DATA/SITAR_1/whi550/ObsData/cubes/H201_abcde_luther_chop.fits" arraysize="62"/>
      <PARAM name="FDR Significance" datatype="float" ucd="stat.param" value="0.1"/>
      <FIELD name="ID" ID="col01" ucd="meta.id" datatype="int" width="5" unit="--"/>
      <FIELD name="Name" ID="col02" ucd="meta.id;meta.main" datatype="char" arraysize="15" unit="--"/>
      <FIELD name="RA" ID="col03" ucd="pos.eq.ra;meta.main" ref="J2000" datatype="float" width="13" precision="6" unit="deg"/>
      <FIELD name="DEC" ID="col04" ucd="pos.galactic.lon;meta.main" ref="J2000" datatype="float" width="13" precision="6" unit="deg"/>
      <FIELD name="w_RA" ID="col05" ucd="phys.angSize;pos.eq.ra" ref="J2000" datatype="float" width="9" precision="2" unit="arcmin"/>
      <FIELD name="w_DEC" ID="col06" ucd="phys.angSize;pos.eq.ra" ref="J2000" datatype="float" width="9" precision="2" unit="arcmin"/>
      <FIELD name="Vel" ID="col07" ucd="phys.veloc;src.dopplerVeloc" datatype="float" width="9" precision="3" unit="km/s"/>
      <FIELD name="w_Vel" ID="col08" ucd="phys.veloc;src.dopplerVeloc;spect.line.width" datatype="float" width="9" precision="3" unit="km/s"/>
      <FIELD name="Integrated_Flux" ID="col09" ucd="phot.flux;spect.line.intensity" datatype="float" width="10" precision="3" unit="Jy.km/s"/>
      <FIELD name="Peak_Flux" ID="col10" ucd="phot.flux;spect.line.intensity" datatype="float" width="10" precision="3" unit="Jy/beam"/>
      <FIELD name="Flag" ID="col11" ucd="meta.code.qual" datatype="char" arraysize="3" unit="--"/>
      <FIELD name="X_Centroid" ID="col12" ucd="pos.cartesian.x" datatype="float" width="7" precision="1" unit=""/>
      <FIELD name="Y_Centroid" ID="col13" ucd="pos.cartesian.y" datatype="float" width="7" precision="1" unit=""/>
      <FIELD name="Z_Centroid" ID="col14" ucd="pos.cartesian.z" datatype="float" width="7" precision="1" unit=""/>
      <FIELD name="X_Av" ID="col15" ucd="pos.cartesian.x" datatype="float" width="6" precision="1" unit=""/>
      <FIELD name="Y_Av" ID="col16" ucd="pos.cartesian.y" datatype="float" width="6" precision="1" unit=""/>
      <FIELD name="Z_Av" ID="col17" ucd="pos.cartesian.z" datatype="float" width="6" precision="1" unit=""/>
      <FIELD name="X_Peak" ID="col18" ucd="pos.cartesian.x" datatype="int" width="7" precision="1" unit=""/>
      <FIELD name="Y_Peak" ID="col19" ucd="pos.cartesian.y" datatype="int" width="7" precision="1" unit=""/>
      <FIELD name="Z_Peak" ID="col20" ucd="pos.cartesian.z" datatype="int" width="7" precision="1" unit=""/>
      <DATA>
        <TABLEDATA>
        <TR>
          <TD>    1</TD><TD> J060925-215712</TD><TD>    92.356613</TD><TD>   -21.953545</TD><TD>    44.51</TD><TD>    39.49</TD><TD>  223.351</TD><TD>   52.765</TD><TD>    14.564</TD>
<TD>     0.213</TD><TD>     </TD><TD> 59.175</TD><TD>140.463</TD><TD>114.439</TD><TD>59.554</TD><TD>140.580</TD><TD>114.536</TD><TD>     59</TD><TD>    140</TD><TD>    114</TD>
        </TR>
... truncated ...
        </TABLEDATA>
      </DATA>
    </TABLE>
  </RESOURCE>
</VOTABLE>
```

F Example Karma Annotation file output

This is the format of the Karma Annotation file, showing the locations of the detected objects. This can be loaded by the plotting tools of the Karma package (for instance, `kvis`) as an overlay on the FITS file.

```
# Duchamp Source Finder results for FITS file:
#  /home/mduchamp/fountain.fits
# Threshold = 4
# No ATrous reconstruction done.
#
COLOR RED
COORD W
CIRCLE 92.3376 -21.9475 0.403992
TEXT 92.3376 -21.9475 1
CIRCLE 91.9676 -26.0193 0.37034
TEXT 91.9676 -26.0193 2
CIRCLE 91.5621 -27.3459 0.437109
TEXT 91.5621 -27.3459 3
CIRCLE 92.8285 -21.6344 0.269914
TEXT 92.8285 -21.6344 4
CIRCLE 90.1381 -28.9838 0.234179
TEXT 90.1381 -28.9838 5
CIRCLE 89.72 -26.6513 0.132743
TEXT 89.72 -26.6513 6
CIRCLE 94.2743 -27.4003 0.195175
TEXT 94.2743 -27.4003 7
CIRCLE 92.2739 -21.6941 0.134538
TEXT 92.2739 -21.6941 8
CIRCLE 89.7133 -25.4259 0.232252
TEXT 89.7133 -25.4259 9
CIRCLE 90.2206 -21.6993 0.266247
TEXT 90.2206 -21.6993 10
CIRCLE 93.8581 -26.5766 0.163153
TEXT 93.8581 -26.5766 11
CIRCLE 91.176 -26.1064 0.234356
TEXT 91.176 -26.1064 12
```

G Robust statistics for a Normal distribution

The Normal, or Gaussian, distribution for mean μ and standard deviation σ can be written as

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/2\sigma^2}.$$

When one has a purely Gaussian signal, it is straightforward to estimate σ by calculating the standard deviation (or rms) of the data. However, if there is a small amount of signal present on top of Gaussian noise, and one wants to estimate the σ for the noise, the presence of the large values from the signal can bias the estimator to higher values.

An alternative way is to use the median (m) and median absolute deviation from the median (s) to estimate μ and σ. The median is the middle of the distribution, defined for a continuous distribution by

$$\int_{-\infty}^{m} f(x)dx = \int_{m}^{\infty} f(x)dx.$$

From symmetry, we quickly see that for the continuous Normal distribution, $m = \mu$. We consider the case henceforth of $\mu = 0$, without loss of generality.

To find s, we find the distribution of the absolute deviation from the median, and then find the median of that distribution. This distribution is given by

$$\begin{aligned} g(x) &= \text{distribution of } |x| \\ &= f(x) + f(-x),\ x \geq 0 \\ &= \sqrt{\frac{2}{\pi\sigma^2}}\, e^{-x^2/2\sigma^2},\ x \geq 0. \end{aligned}$$

So, the median absolute deviation from the median, s, is given by

$$\int_{0}^{s} g(x)dx = \int_{s}^{\infty} g(x)dx.$$

If we use the identity

$$\int_{0}^{\infty} e^{-x^2/2\sigma^2}dx = \sqrt{\pi\sigma^2/2}$$

we find that

$$\int_{s}^{\infty} e^{-x^2/2\sigma^2}dx = \sqrt{\pi\sigma^2/2} - \int_{0}^{s} e^{-x^2/2\sigma^2}dx.$$

Hence, to find s we simply solve the following equation (setting $\sigma = 1$ for

simplicity – equivalent to stating x and s in units of σ):

$$\int_0^s e^{-x^2/2} dx - \sqrt{\pi/8} = 0.$$

This is hard to solve analytically (no nice analytic solution exists for the finite integral that I'm aware of), but straightforward to solve numerically, yielding the value of $s = 0.6744888$. Thus, to estimate σ for a Normally distributed data set, one can calculate s, then divide by 0.6744888 (or multiply by 1.4826042) to obtain the correct estimator.

Note that this is different to solutions quoted elsewhere, specifically in Meyer et al. (2004), where the same robust estimator is used but with an incorrect conversion to standard deviation – they assume $\sigma = s\sqrt{\pi/2}$. This, in fact, is the conversion used to convert the *mean* absolute deviation from the mean to the standard deviation. This means that the cube noise in the HIPASS catalogue (their parameter $\mathrm{Rms}_{\mathrm{cube}}$) should be 18% larger than quoted.

H How Gaussian noise changes with wavelet scale

The key element in the wavelet reconstruction of an array is the thresholding of the individual wavelet coefficient arrays. This is usually done by choosing a level to be some number of standard deviations above the mean value.

However, since the wavelet arrays are produced by convolving the input array by an increasingly large filter, the pixels in the coefficient arrays become increasingly correlated as the scale of the filter increases. This results in the measured standard deviation from a given coefficient array decreasing with increasing scale. To calculate this, we need to take into account how many other pixels each pixel in the convolved array depends on.

To demonstrate, suppose we have a 1-D array with N pixel values given by $F_i,\ i = 1, ..., N$, and we convolve it with the B_3-spline filter, defined by the set of coefficients $\{1/16, 1/4, 3/8, 1/4, 1/16\}$. The flux of the ith pixel in the convolved array will be

$$F'_i = \frac{1}{16}F_{i-2} + \frac{1}{4}F_{i-1} + \frac{3}{8}F_i + \frac{1}{4}F_{i+1} + \frac{1}{16}F_{i+2}$$

and the flux of the corresponding pixel in the wavelet array will be

$$W'_i = F_i - F'_i = \frac{-1}{16}F_{i-2} - \frac{1}{4}F_{i-1} + \frac{5}{8}F_i - \frac{1}{4}F_{i+1} - \frac{1}{16}F_{i+2}$$

Now, assuming each pixel has the same standard deviation $\sigma_i = \sigma$, we can work out the standard deviation for the wavelet array:

$$\sigma'_i = \sigma\sqrt{\left(\frac{1}{16}\right)^2 + \left(\frac{1}{4}\right)^2 + \left(\frac{5}{8}\right)^2 + \left(\frac{1}{4}\right)^2 + \left(\frac{1}{16}\right)^2} = 0.72349\ \sigma$$

Thus, the first scale wavelet coefficient array will have a standard deviation of 72.3% of the input array. This procedure can be followed to calculate the necessary values for all scales, dimensions and filters used by *Duchamp*.

Calculating these values is clearly a critical step in performing the reconstruction. The method used by Starck & Murtagh (2002) was to simulate data sets with Gaussian noise, take the wavelet transform, and measure the value of σ for each scale. We take a different approach, by calculating the scaling factors directly from the filter coefficients by taking the wavelet transform of an array made up of a 1 in the central pixel and 0s everywhere else. The scaling value is then derived by taking the square root of the sum (in quadrature) of all the wavelet coefficient values at each scale. We give the scaling factors for the three filters available to *Duchamp* below. These values are hard-coded into *Duchamp*, so no on-the-fly calculation of them is necessary.

Memory limitations prevent us from calculating factors for large scales, particularly for the three-dimensional case (hence the – symbols in the ta-

bles). To calculate factors for higher scales than those available, we divide the previous scale's factor by either $\sqrt{2}$, 2, or $\sqrt{8}$ for 1D, 2D and 3D respectively.

- **B_3-Spline Function:** $\{1/16, 1/4, 3/8, 1/4, 1/16\}$

Scale	1 dimension	2 dimension	3 dimension
1	0.723489806	0.890796310	0.956543592
2	0.285450405	0.200663851	0.120336499
3	0.177947535	0.0855075048	0.0349500154
4	0.122223156	0.0412474444	0.0118164242
5	0.0858113122	0.0204249666	0.00413233507
6	0.0605703043	0.0101897592	0.00145703714
7	0.0428107206	0.00509204670	0.000514791120
8	0.0302684024	0.00254566946	–
9	0.0214024008	0.00127279050	–
10	0.0151336781	0.000636389722	–
11	0.0107011079	0.000318194170	–
12	0.00756682272	–	–
13	0.00535055108	–	–

- **Triangle Function:** $\{1/4, 1/2, 1/4\}$

Scale	1 dimension	2 dimension	3 dimension
1	0.612372436	0.800390530	0.895954449
2	0.330718914	0.272878894	0.192033014
3	0.211947812	0.119779282	0.0576484078
4	0.145740298	0.0577664785	0.0194912393
5	0.102310944	0.0286163283	0.00681278387
6	0.0722128185	0.0142747506	0.00240175885
7	0.0510388224	0.00713319703	0.000848538128
8	0.0360857673	0.00356607618	0.000299949455
9	0.0255157615	0.00178297280	–
10	0.0180422389	0.000891478237	–
11	0.0127577667	0.000445738098	–
12	0.00902109930	0.000222868922	–
13	0.00637887978	–	–

- **Haar Wavelet:** $\{0, 1/2, 1/2\}$

Scale	1 dimension	2 dimension	3 dimension
1	0.707167810	0.433012702	0.935414347
2	0.500000000	0.216506351	0.330718914
3	0.353553391	0.108253175	0.116926793
4	0.250000000	0.0541265877	0.0413398642
5	0.176776695	0.0270632939	0.0146158492
6	0.125000000	0.0135316469	0.00516748303