Cluster Statistics and Coincidences

Zun Zar Chi Naing
Simmons College

July 2013
University of Florida
International REU 2013

Università "La Sapienza" and Istituto Nazionale di Fisica Nucleare,sezione di Roma
Mentor: Sergio Frasca

Abstract

Currently, there are about 100 million candidates found in each of the two different time periods by the Virgo antenna, whose numbers could be greatly enhanced with time. Each candidate source is defined by a position in the sky, a frequency and a spin-down parameter. The main reason behind this big number is that there are disturbances mixed with the candidate sources. By means of clustering the candidates of similar parameters, these disturbances can be filtered out. Matlab scripts have been written to compare the clusters based on different parameters at different time periods. Histograms of the comparisons have been obtained and observations of different patterns in different frequency bands of different time periods have been made. However, difficulties remain in areas such as effective algorithms in distinguishing the true sources and disturbances, and evaluation of the significance of the clustering histograms. With further understanding of the clusters behavior, work on coincidences of clusters has to be carried out. This work is currently under progress and effective codes for coincidences have yet to be written since running time of each script could range from days to weeks.

Introduction

Periodic gravitational waves, in the range of frequency of the Virgo antenna, are produced by non-axisymmetric rotating neutron stars. The periodic sources are the only type of gravitational signal that can be detected by a single gravitational antenna with certainty since the estimation of the source parameters can be done with the highest precision. Hierarchical procedures, based on alternating incoherent and coherent steps, were introduced to overcome the problem of the massive computing power needed for the optimal detection of periodic sources. The first incoherent step consists of Hough transform based on the collection of short FFT periodograms, which produces candidates of possible sources. Then, with a coherent step, the candidates* are analyzed and the search is refined again, which is followed by a new incoherent step until the full sensitivity is reached. [1] After a series of filtering, only "clean" spectrum is left. These "clean" sources are the ones that have been analyzed in the current work.

* Candidates are the potential sources that produce gravitational waves.

Since a periodic source is permanent, one can repeat the experiment with the same or better antenna and check the ''true existence" of a source candidate. This is what is called "coincidences" between candidates in different periods. The probability to have a coincidence between the two sets of candidates in two 4-month periods is of the order of 10^{-20}. [1] When dealing with the clusters with a big number of candidates in them, it takes longer to process the coincidences, and most likely, those are filled with disturbances. Therefore, preferentially, only clusters with smaller candidates are chosen and coincidence is done between those clusters in two different time periods.

Clustering

"Clean" candidates, after filtering and considering the earth rotation, were grouped together into groups called clusters. Candidates are put into the same cluster if they are in the same range of parameters. Parameters in consideration include frequency, amplitude, spin-down, lamda and beta (longitude and latitude of sky location), and numerosity (number of candidates in a group). These clusters are grouped under a variety of frequency bandwidth, starting from 10-30 Hz, then 30-40Hz, up to 120-130Hz. Since the candidates are collected in two different time periods, there are always two clusters of the same frequency bandwidth but at two different time periods, namely VSR2 and VSR4. All these clusters are labeled as "VSR2_010030_cl" and "VSR4_010030_cl" - VSR2 and VSR4 being two different time period data collection, 010030 being frequency bandwidth of 10-30Hz, and 'cl' being clusters. This work has already been done before my arrival at La Sapienza. Table 1 shows an observation of number of clusters in different frequency range for two different time periods. For VSR2, the number of clusters seems to increase as the frequency increases until the frequency reaches 110Hz, then the number of clusters to starts to go down. However, it is not the same case for VSR4. For VSR4, the number of clusters just randomly increases and decreases across the different frequency ranges. The amplitude and the frequency of the waves change with time because the emission of gravitational radiation determines an evolution of the emitting source (due to the gravitational radiation reaction). This change in amplitude and the frequency over time contributes to this variation of the number of clusters between the two different time periods.

Coincidences

After candidates with close parameters are grouped together into clusters in each time period, the two groups of clusters from VSR2 and VSR4 are compared. When the clusters from VSR2 and VSR4 match in their given parameters, they are again grouped into coincidences. Similar to creating clusters, when creating coincidences, only clusters with 10 or less candidates are compared for more accuracy and time efficiency. When coincidences are obtained, the candidates in that group of coincidences are again analyzed more carefully and these candidates are filtered again through coherent and incoherent steps.

Fr range (Hz)	VSR2 (No: of clusters)	VSR4 (No: of clusters)
10-30	101,210	79,699
30-40	187,520	110,001
40-50	211,209	150,189
50-60	337,050	265,873
60-70	419,547	254,405
70-80	492,152	332,206
80-90	326,139	233,269
90-100	273,189	201,429
100-110	174,340	256,332
110-120	83,386	192,115
120-130	40,157	61,333

Table 1: Comparisons of number of clusters in different frequency ranges across two different time periods

Matlab

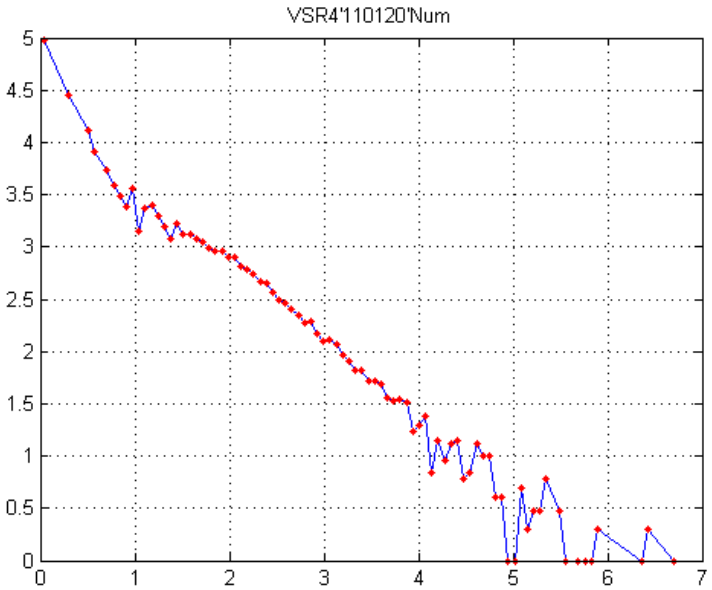
Analysis of the clusters plays a very important part in creating coincidences. If the clusters were created incorrectly, then creating coincidences based on those clusters would be a waste of time. In analyzing the clusters, Matlab programming is used. Matlab programming is very effective in this situation because all the clusters are a structure with a number of parameter categories. The clusters were analyzed by creating histograms of different parameters. Different histograms of different parameters are created based on certain parameter. To do so, first we will have to see the histograms of candidates at each parameter. From it, if we see two or more different patterns going on in all of the parameter histogram, we will pick a point where the patterns differ. Then, we can pick any parameter and based on that parameter, all the different parameters are compared again, i.e., if the parameter, numerosity, is selected to be the main parameter based on which the candidates are compared, we pick a point in the numerosity histogram in which the patterns of candidate behavior differ. As shown in Table 2, a point where the patterns differ, numerosity 10, is picked. Then, the behaviors of all parameters are compared in the same histograms when the numerosity is less than 10 and when the numerosity is greater than 10. The main intention is to see how the clusters behave when just viewed under certain parameter.

Before creating histograms, an effective Matlab script has to be written so that clusters can be created correctly depending on their given parameter, which is one of the main difficulties. A lot of errors come up as the script is being written. Mostly, the difficulties reside in setting and determining the parameters. There were a lot of struggles along the way of improvising the scripts as sometimes each script would take days to weeks to be operated.

Initially, scripts were written just to see what the clusters look like on histograms. Matlab scripts to produce bar histograms were written initially but since nothing can be interpreted out of the bar histograms, Matlab script was edited so that it will produce log based histograms. The scripts were initially written just based on numerosity (number of candidates in a cluster). Histograms of different parameters were obtained based on numerosity, and it was observed that there were two different families in the data, meaning there were two or more different slopes in each parameter. Using the existing program in Matlab written by Sergio Frasca, slopes of different

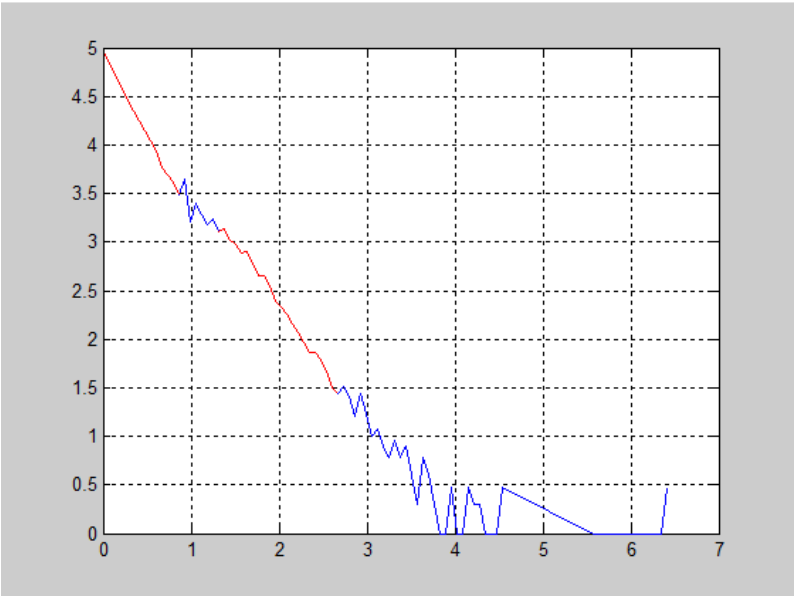
families were obtained. With this observation, further more scripts were written to analyze the two different families. Prototype of Matlab script can be found in the appendix.

Data Analysis and Results



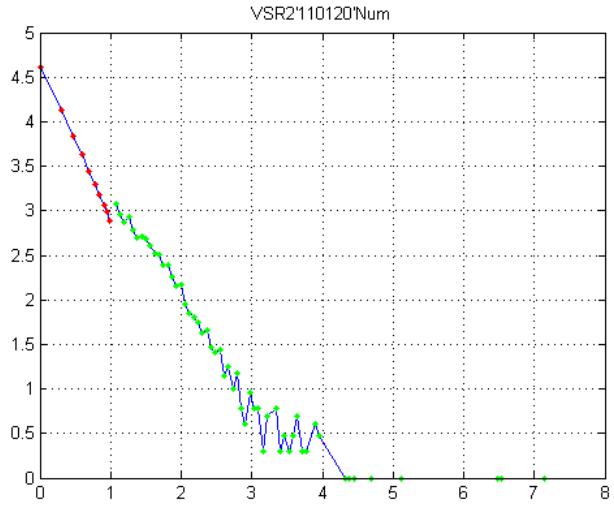
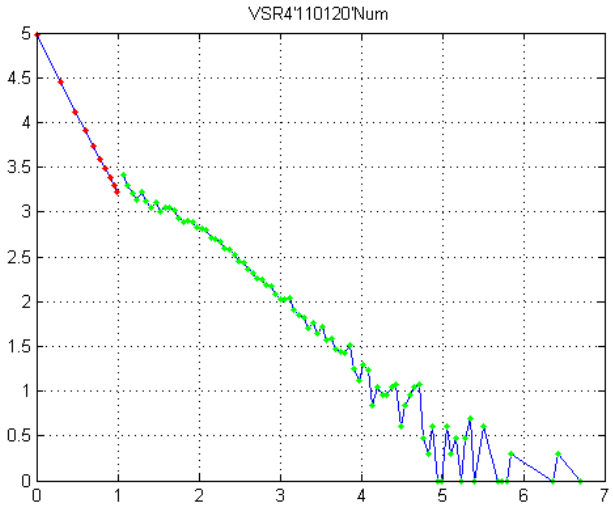
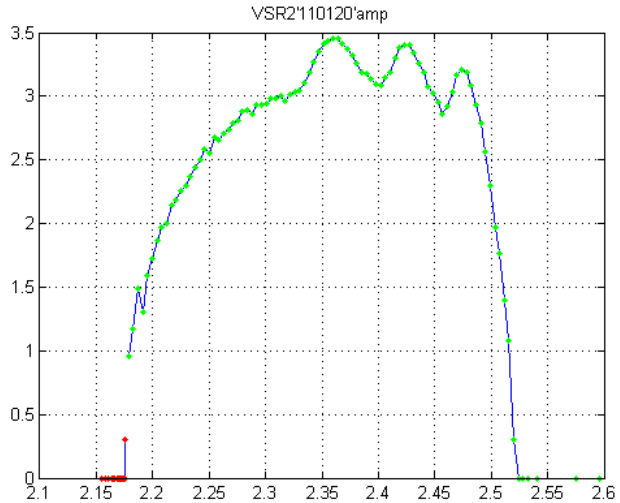
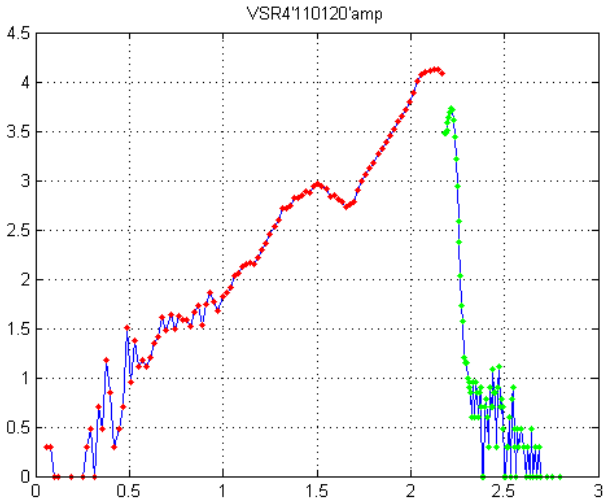
Graph 1: Log histogram of VSR4_110120_cl based on number of candidates it obtained.

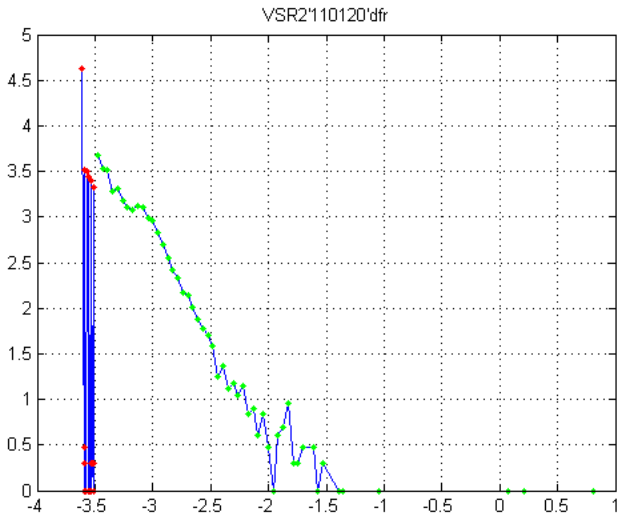
From this histogram as shown in Graph 1, we can see that there are two different families that have different slopes. The mixture of the true sources and the disturbances can create different spread of slopes. To understand better, the slopes of different families are obtained as in Graph 2.

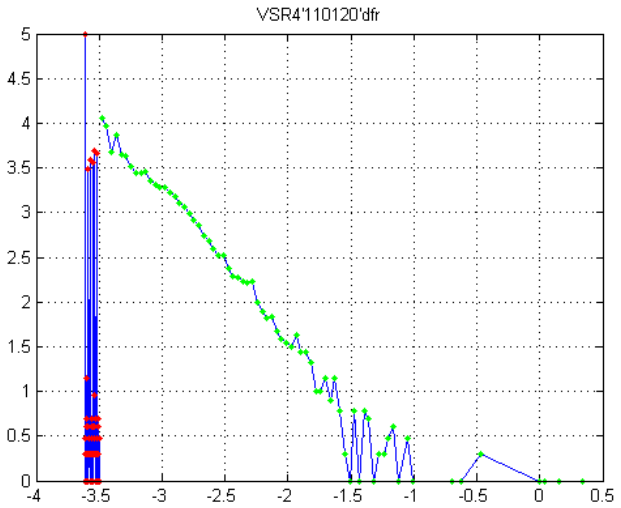


Graph 2: Showing the different slopes of two different families in a cluster. The red lines are the regions where the slopes are obtained. From the left to right, slope (1) = -1.7155, slope (2)=-1.2773

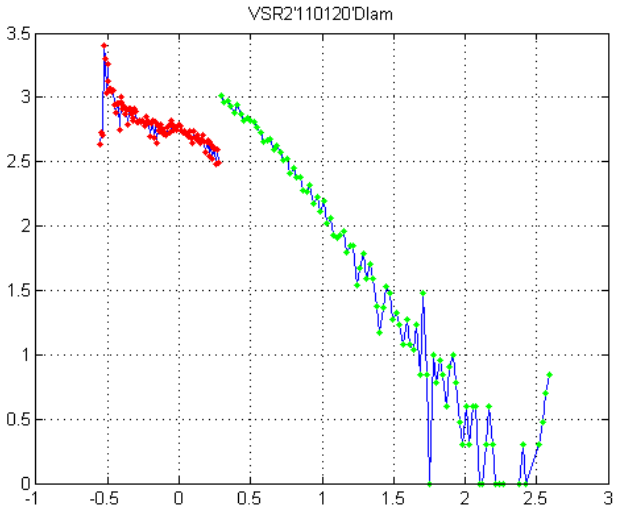
Looking at a whole histogram, I chose a point where there is a shift in the histogram. For example, in the graph 2, the first ten clusters act in a same behavior, all going down slope. After that first ten clusters, the cluster behavior shifts. Therefore, I would take number of clusters, 10 as the point where I would observe the differences of the cluster pattern before numerosity 10 and after numerosity 10. This is done the same way in other parameters as well. It is important to use the same number when comparing the two clusters of same frequency range in two different time periods. The table below shows an excerpt of the cluster comparisons at different time periods based on different parameter.

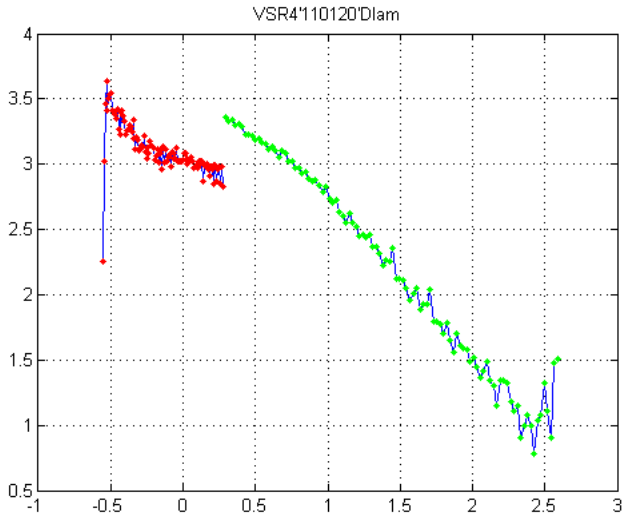
VSR2	VSR4
Comparing clusters based on different numerocity (red being num <=10, green being num > 10)	
	
Comparing clusters based on different amplitude (red being amp <=150, green being amp > 150)	
	
Comparing clusters based on freq diff (red being dfr <=0.000316, green being dfr > 0.000316)	



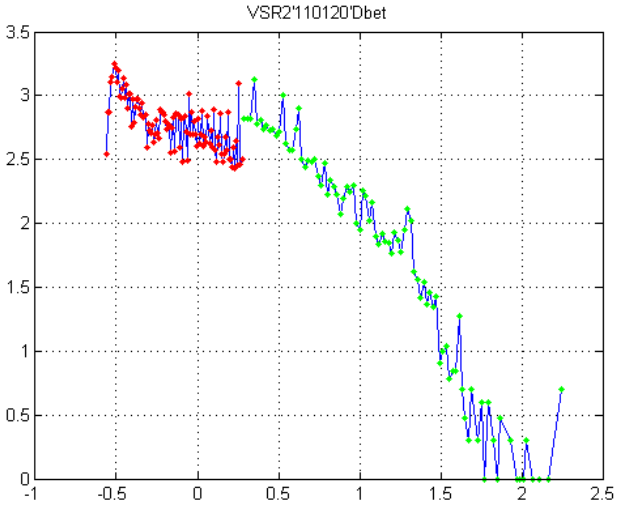


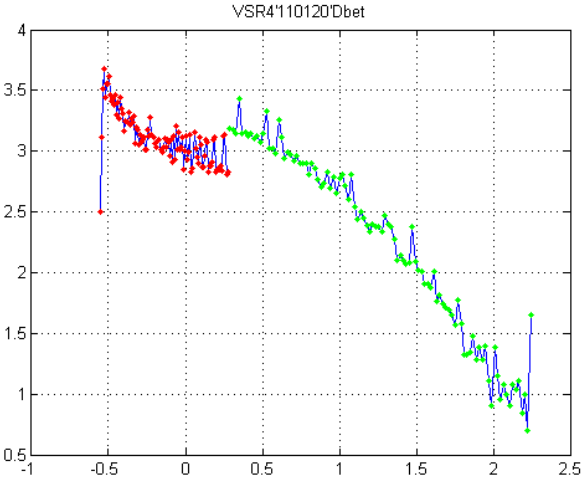
Comparing clusters based on different lamda diff (red being dlam <=1.90, green being dlam > 1.90)





Comparing clusters based on different beta diff (red being dbet <=1.90, green being dbet > 1.90)





Comparing clusters based on spindown diff (red being dsd <=4.467e-10, green being dsd > 4.467e-10)

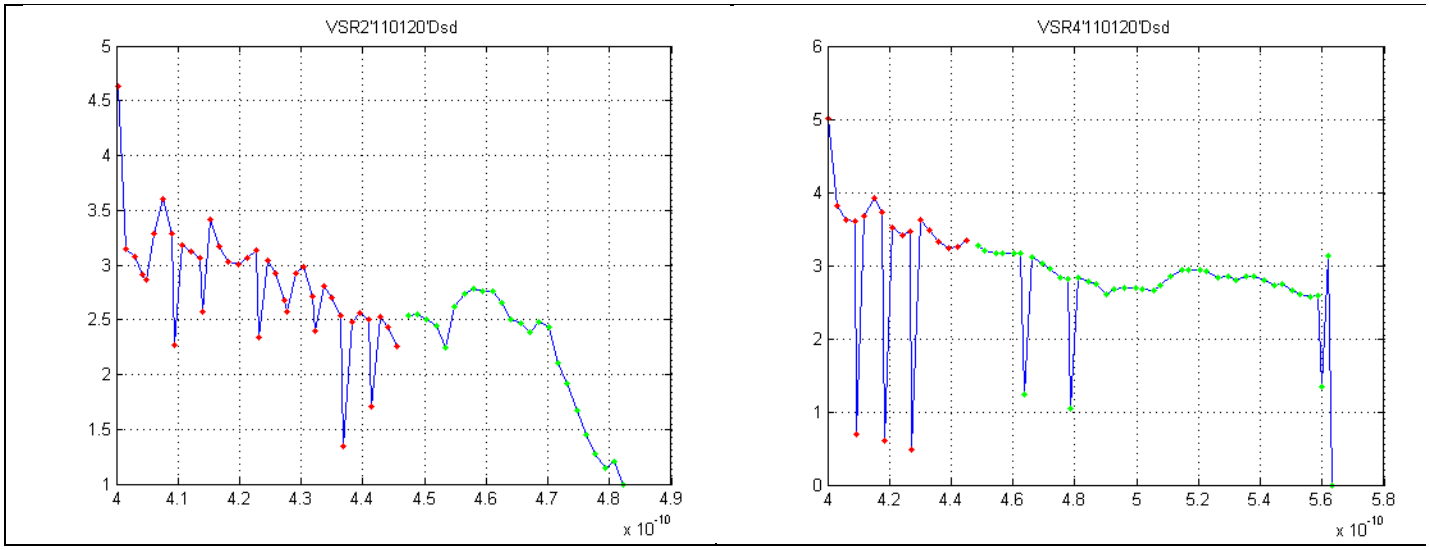


Table 2: Histograms of clusters of different parameter comparisons based on different parameter at two different time periods

From the above table, we could clearly see the different patterns the two clusters from different time periods would form even though they are at the same frequency ranges. Ideally, since the sources that are analyzed are periodic, the pattern that the histograms produce would be similar in both different time periods. Since this is not the case here, it could be explained by the shift in frequency and amplitude with time. In all of these histograms, two different families are compared in one graph to see these significant differences between them. As the number of candidates increases in clusters, the area underneath the curve becomes bigger; this most likely implies that those clusters are more filled with disturbances.

Coincidences Statistics

Although the work on coincidences is yet still in progress, the results so far we have gathered has been interesting, and definitely give road to more work, such as analysis and understanding of what the coincidences mean. With the current Matlab script to create coincidences between VSR2 and VSR4 is run, the following histograms are obtained.

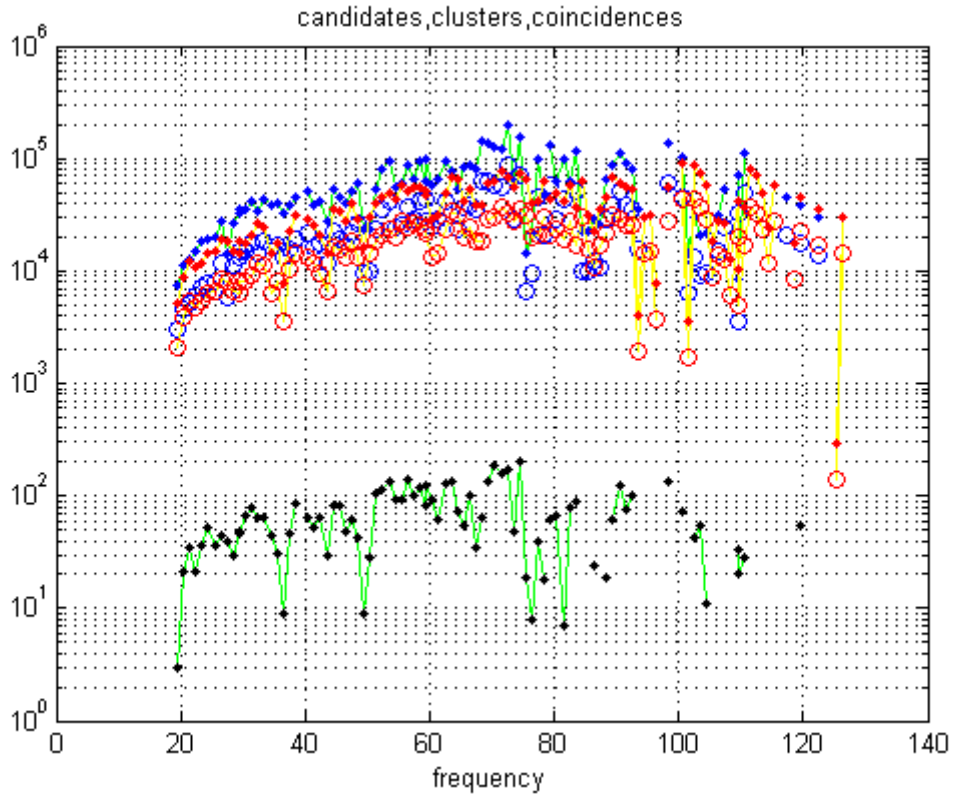


Figure 1: Spread of number of candidates, clusters and coincidences of VSR2 and VSR4 over a range of frequency.

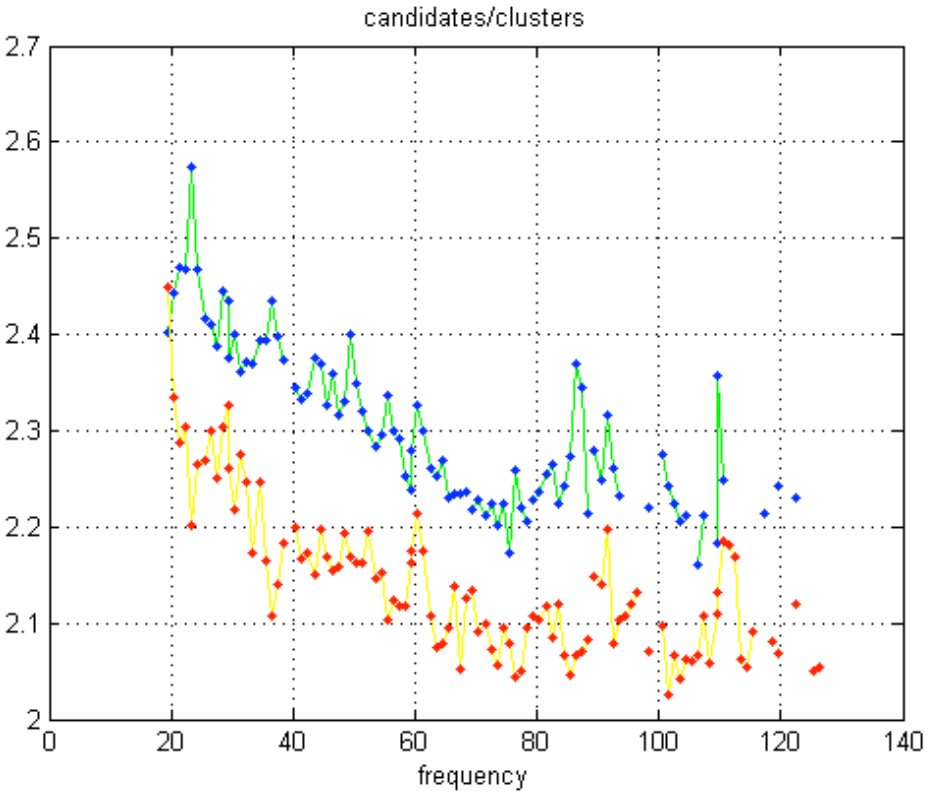


Figure 2: The histogram of the ratio of the number of candidates over number of clusters for both VSR2 and VSR4 over a range of frequency

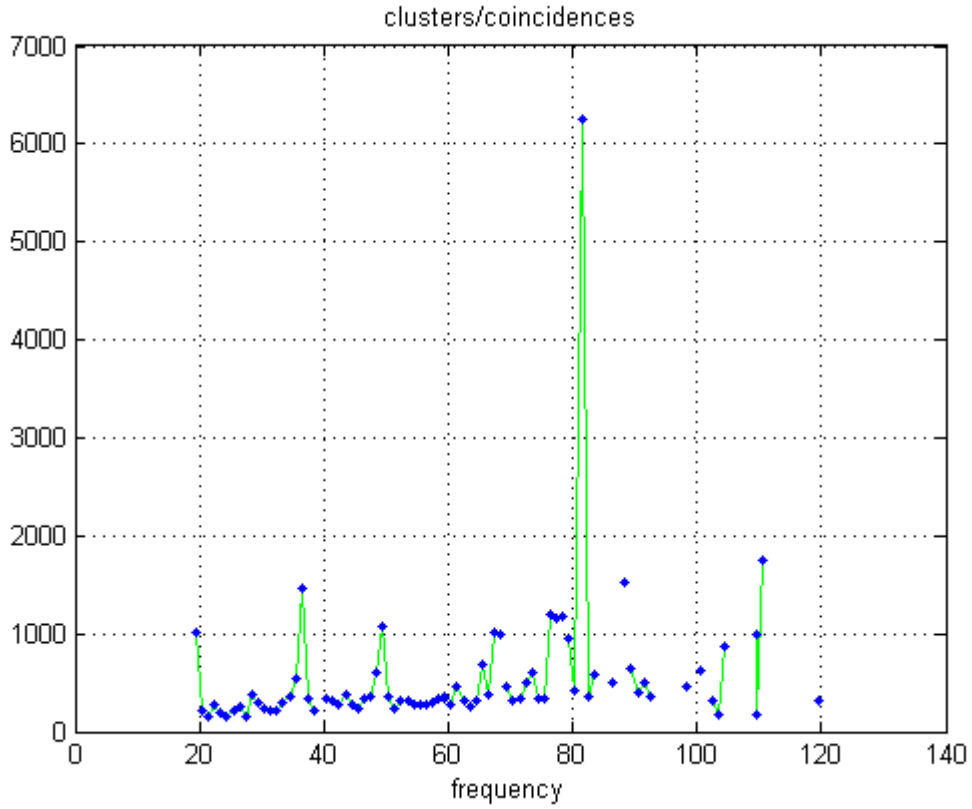


Figure 3: The histogram of the ratio of the number of clusters over number of coincidences over a range of frequency

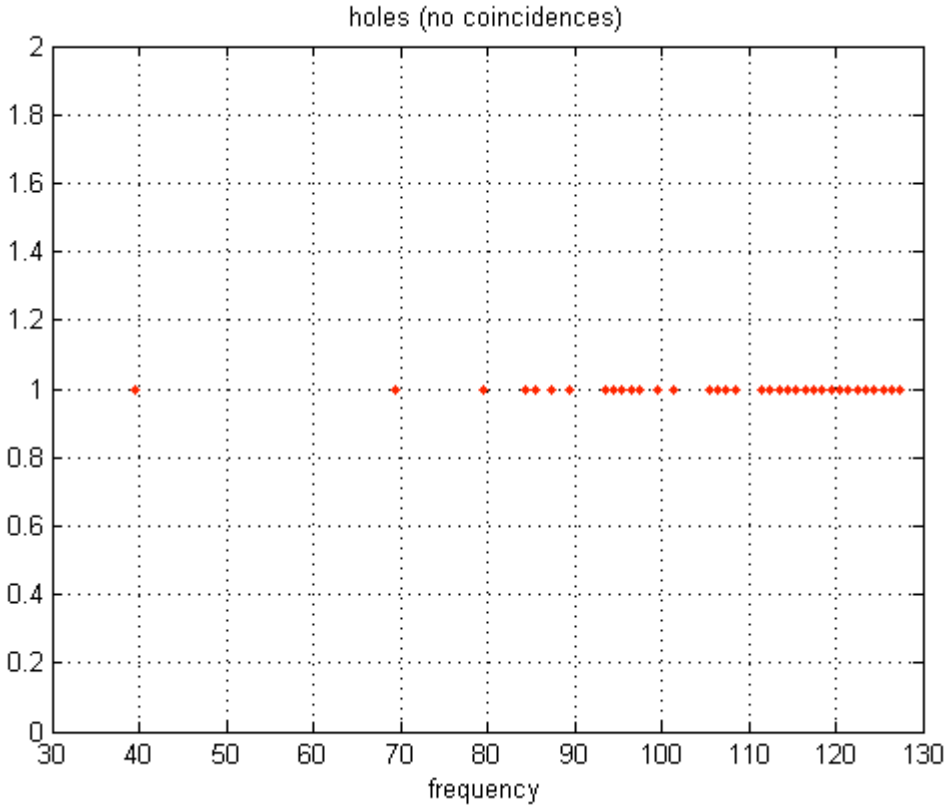


Figure 4: The histogram showing number of holes over a range of frequency

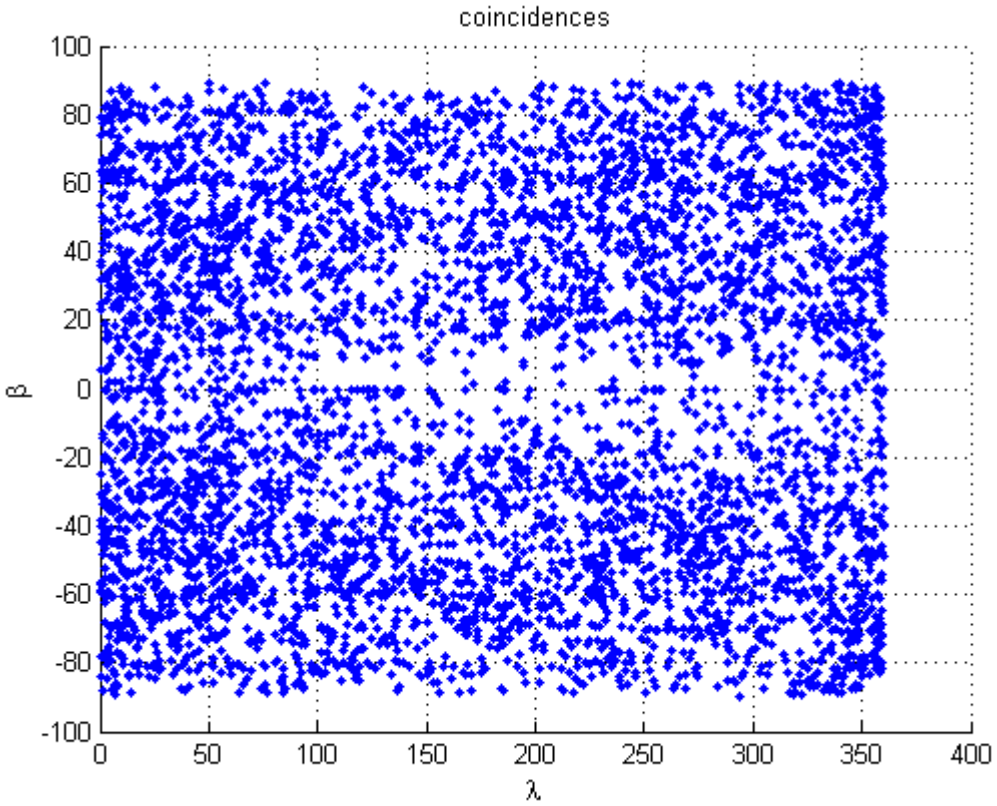


Figure 5: Spread of coincidences in a sky map

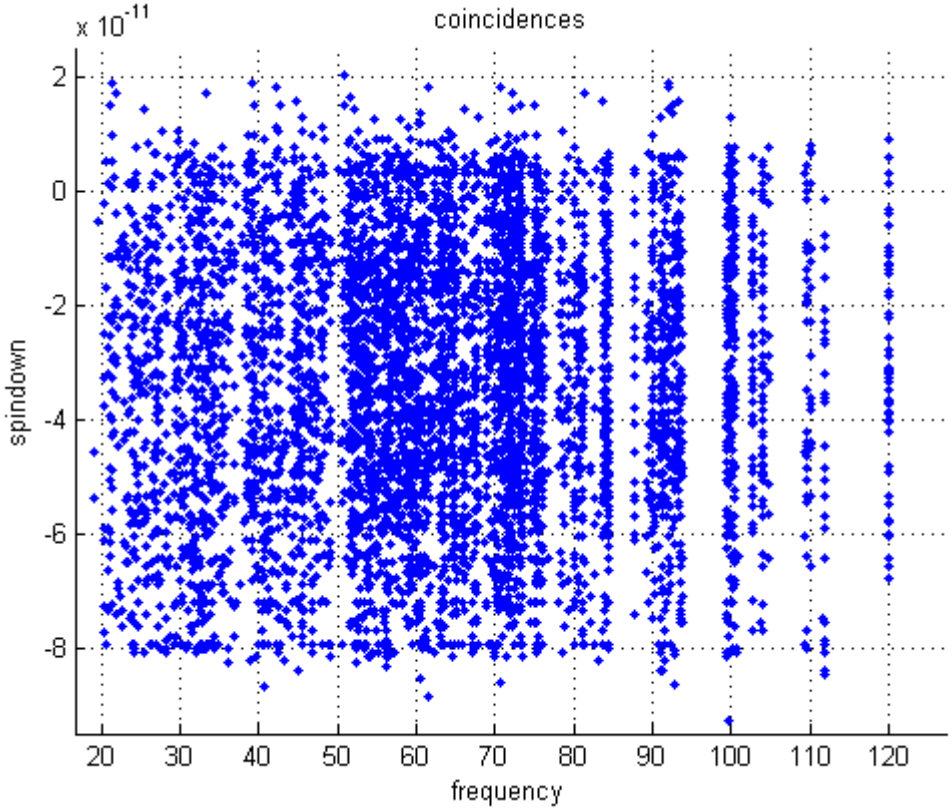


Figure 6: Spread of coincidences over frequency and spin-down

Figure 1 shows the spread of number of candidates, clusters and coincidences of VSR2 and VSR4 over a range of frequency. Although it is hard to see in this figure, when zoomed in, the blue dots and the red dots represent the number of candidates in VSR2 and VSR4 respectively. The blue and red circles represent the number of clusters in VSR2 and VSR4 respectively and the lowermost black dots represent the number of coincidences between VSR2 and VSR4. As expected, the number of candidates would be highest, followed by the number of clusters and coincidences. Figure 2 and figure 3 show the representation of the ratio between number of candidates and clusters and the ratio of between number of clusters and coincidences respectively. These two figures give us a general idea of how the clusters are grouped together over a range of frequency. As shown in all of the figures above, there are some regions where there are no coincidences. These regions are called holes. The presence of the number of holes over a range of frequency is shown in figure 4. The existence of the holes means that there are some errors when creating the clusters, or in determining the parameters. Further work has to be done to understand more on holes. There are currently about 5,000 coincidences and when these coincidences are placed according the location in the sky, we get the type of histogram as shown in figure 5. We could see the arrangement of coincidences in all different parameters, and in figure 6 we could see the spread of the coincidences with respect to frequency and spin-down.

Conclusion

Matlab codes to create clusters and to analyze clusters have been written. Although no definite conclusions have been made based on these histograms, interesting patterns of cluster behavior is observed, and could definitely conclude that there are still many disturbances mixed in the candidates. Further work on coincidences is being carried out to further understand the nature of true candidates and to differentiate them from disturbances. Since data analysis group at Rome works with hierarchical procedures, after the step of obtaining coincidences, each candidate will be analyzed again and a new coincidence will be obtained. All these coincidences will be analyzed with much more precaution and understanding.

Acknowledgment

I would like to thank the University of Florida International REU program and National Science Foundation (NSF) for giving me an opportunity to experience research environment in Rome, Italy. Many thanks go to my mentor, Sergio Frasca, for patiently guiding me and explaining me throughout the program as I started out as a beginner in Matlab. Additionally, I would like to thank Pia Astone for arranging all the necessary things for my arrival in Rome. I also would like to appreciate all the people at the residence who have made my stay in Italy memorable.

References

1. *All sky search for continuous gravitational waves*. S.Frasca. *Warsaw, International Banach Center, 1st – 10th September 2003.*

2. Periodic Sources of Gravitational Radiation, Sept. 2000.

Appendix

A prototype of matlab script to create cluster and to create histograms at different parameter

```
% trial2_amp
% Comparing clusters based of amplitude

len=length(VSR4_110120_cl_clust2);
numer=zeros(1,len);

for i = 1:len
    numer(i)=VSR4_110120_cl_clust2(i).ampmax;
end

ind=find(numer <= 150);

cl=VSR4_110120_cl_clust2(ind);

Num=length(ind);
Fr=zeros(1,Num);
Dfr=Fr;
lam=Fr;
Dlam=Fr;
bet=Fr;
Dbet=Fr;
sd=Fr;
Dsd=Fr;
num=Fr;
Dbet1=Fr;
Dlam1=Fr;
amp=Fr;

for i = 1:Num
    Fr(i)=cl(i).fr;
    Dfr(i)=cl(i).frmax-cl(i).frmin;
    lam(i)=cl(i).lam;
    Dlam(i)=cl(i).lammax-cl(i).lammin;
    Dlam1(i)=Dlam(i)./cl(i).dlam;
    bet(i)=cl(i).bet;
    Dbet(i)=cl(i).betmax-cl(i).betmin;
    Dbet1(i)=Dbet(i)./cl(i).dbet;
    sd(i)=cl(i).sd;
    Dsd(i)=cl(i).sdmax-cl(i).sdmin;
    num(i)=cl(i).num;
    amp(i)=cl(i).ampmax;
end

ind1=find(numer > 150);

cl1=VSR4_110120_cl_clust2(ind1);

Numx=length(ind1);
Frx=zeros(1,Numx);
Dfrx=Frx;
lamx=Frx;
Dlamx=Frx;
betx=Frx;
```

```
Dbetx=Frx;
sdx=Frx;
Dsdx=Frx;
numx=Frx;
Dbet1x=Frx;
Dlam1x=Frx;
ampx=Frx;

for b = 1:Numx
    Frx(b)=cl1(b).fr;
    Dfrx(b)=cl1(b).frmax-cl1(b).frmin;
    lamx(b)=cl1(b).lam;
    Dlamx(b)=cl1(b).lammax-cl1(b).lammin;
    Dlam1x(b)=Dlamx(b)./cl1(b).dlam;
    betx(b)=cl1(b).bet;
    Dbetx(b)=cl1(b).betmax-cl1(b).betmin;
    Dbet1x(b)=Dbetx(b)./cl1(b).dbet;
    sdx(b)=cl1(b).sd;
    Dsdx(b)=cl1(b).sdmax-cl1(b).sdmin;
    numx(b)=cl1(b).num;
    ampx(b)=cl1(b).ampmax;
end

[h X]=hist(log10(num),100);
H=log10(h);
ii=find(isfinite(H));
Hnum=H(ii);
Xnum=X(ii);
figure,plot(Xnum,Hnum),grid on,hold on,plot(Xnum,Hnum,'r.'), hold all
[a K]=hist(log10(numx),100);
A=log10(a);
iii=find(isfinite(A));
Anum=A(iii);
Knum=K(iii);
plot (Knum,Anum),hold on, plot (Knum,Anum,'g.'),hold off
title('VSR4''110120''Num')

[h X]=hist(log10(amp),100);
H=log10(h);
ii=find(isfinite(H));
Hamp=H(ii);
Xamp=X(ii);
figure,plot(Xamp,Hamp),grid on,hold on,plot(Xamp,Hamp,'r.'), hold all
[a K]=hist(log10(ampx),100);
A=log10(a);
iii=find(isfinite(A));
Aamp=A(iii);
Kamp=K(iii);
plot (Kamp,Aamp),hold on, plot (Kamp,Aamp,'g.'),hold off
title('VSR4''110120''amp')

[h X]=hist(log10(Dfr),100);
H=log10(h);
ii=find(isfinite(H));
Hdfr=H(ii);
Xdfr=X(ii);
figure,plot(Xdfr,Hdfr),grid on,hold on,plot(Xdfr,Hdfr,'r.'), hold all
[a K]=hist(log10(Dfrx),100);
A=log10(a);
iii=find(isfinite(A));
Adfr=A(iii);
Kdfr=K(iii);
```

```
plot (Kdfr,Adfr),hold on, plot (Kdfr,Adfr,'g.'),hold off
title('VSR4''110120''dfr')

[h X]=hist(log10(Dlam),100);
H=log10(h);
ii=find(isfinite(H));
Hdlam=H(ii);
Xdlam=X(ii);
figure,plot(Xdlam,Hdlam),grid on,hold on,plot(Xdlam,Hdlam,'r.'), hold all
[a K]=hist(log10(Dlamx),100);
A=log10(a);
iii=find(isfinite(A));
Adlam=A(iii);
Kdlam=K(iii);
plot (Kdlam,Adlam),hold on, plot (Kdlam,Adlam,'g.'),hold off
title('VSR4''110120''Dlam')

[h X]=hist(log10(Dlam1),100);
H=log10(h);
ii=find(isfinite(H));
Hdlam1=H(ii);
Xdlam1=X(ii);
figure,plot(Xdlam1,Hdlam1),grid on,hold on,plot(Xdlam1,Hdlam1,'r.'),hold all
[a K]=hist(log10(Dlam1x),100);
A=log10(a);
iii=find(isfinite(A));
Adlam1=A(iii);
Kdlam1=K(iii);
plot (Kdlam1,Adlam1),hold on, plot (Kdlam1,Adlam1,'g.'),hold off
title('VSR4''110120''Dlam1')

[h X]=hist(log10(Dbet),100);
H=log10(h);
ii=find(isfinite(H));
Hdbet=H(ii);
Xdbet=X(ii);
figure,plot(Xdbet,Hdbet),grid on,hold on,plot(Xdbet,Hdbet,'r.'),hold all
[a K]=hist(log10(Dbetx),100);
A=log10(a);
iii=find(isfinite(A));
Adbet=A(iii);
Kdbet=K(iii);
plot (Kdbet,Adbet),hold on, plot (Kdbet,Adbet,'g.'),hold off
title('VSR4''110120''Dbet')

[h X]=hist(log10(Dbet1),100);
H=log10(h);
ii=find(isfinite(H));
Hdbet1=H(ii);
Xdbet1=X(ii);
figure,plot(Xdbet1,Hdbet1),grid on,hold on,plot(Xdbet1,Hdbet1,'r.'),hold all
[a K]=hist(log10(Dbet1x),100);
A=log10(a);
iii=find(isfinite(A));
Adbet1=A(iii);
Kdbet1=K(iii);
plot (Kdbet1,Adbet1),hold on, plot (Kdbet1,Adbet1,'g.'),hold off
title('VSR4''110120''Dbet1')

[h X]=hist(Dsd,100);
H=log10(h);
ii=find(isfinite(H));
```

```
Hdsd=H(ii);
Xdsd=X(ii);
figure,plot(Xdsd,Hdsd),grid on,hold on,plot(Xdsd,Hdsd,'r.'),hold all
[a K]=hist(Dsdx,100);
A=log10(a);
iii=find(isfinite(A));
Adsd=A(iii);
Kdsd=K(iii);
plot(Kdsd,Adsd),grid on,hold on,plot(Kdsd,Adsd,'g.')
title('VSR4''110120''Dsd')
```