Statistical analysis of narrow-band signals at setilive.org

Igor Nikitin
Department of High Performance Analytics
Fraunhofer Institute for Algorithms and Scientific Computing
Schloss Birlinghoven, 53757 Sankt Augustin, Germany

igor.nikitin@scai.fraunhofer.de

Abstract. SETILive is a web project forwarding radio signals from SETI Institute's Allen Telescope Array (ATA) for the analysis of volunteers. It contains a large archive with more than 1.5 millions observations for more than 7.5 thousands observation targets, including directions to exoplanets discovered by telescope Kepler and other sources. It also supports various tools for signal collection and classification. Till recent time it supported live feeds of signals from ATA together with a feedback loop, a possibility to interrupt the schedule and repeat the observation of an interesting signal registered by sufficiently many viewers. Unfortunately, since 12-Oct-2014 the live feeds have been discontinued. We hope that the project will persist, taking into account the importance of the search subject, the worldwide interest to the topic and the value of already collected data. In this paper we present the results of statistical analysis of data stored in SETILive archive, using Radon transform and specially constructed filter for selection of single beams, potential signals of ET origination. We will also estimate statistical significance of signals depending on their signal-to-noise ratio using Monte Carlo simulation and select 28 strong signals and totally 1072 statistically significant signals in the archive.

Fig.1. Synthetic signal of different strength and its Radon transform.
Introduction

Signal analysis at SETI [1] uses so called waterfall plots, representing a moving window Fourier decomposition of the signal with frequency taken as horizontal axis and time as vertical axis. Especially interesting features on the waterfall plots are straight lines, corresponding to narrow signals with slowly drifting frequency. Given a source signal of a fixed frequency, Doppler effect will produce a frequency shift, while the rotation and orbital motion of the Earth and similar motions of the signal source will make this shift varying in time. If the observation period is much less than the periods of these motions, the drift will be approximately linear in time and will produce straight lines on the waterfall plots. Earlier it has been proposed to use Radon transform for searching straight lines on the waterfall plots [2,3]. Radon transform performs an integration of a signal along straight lines in various directions, accumulating lines to points. In this way signal-to-noise ratio (SNR) can be increased. An example on Fig.1 shows a synthetic signal of the form

\[ s(f,t) = SNR \exp(-((f-f_0-kt)/b)^2) + N(f,t) \]

where \( f \) denotes frequency, \( t \) – time, \( f_0+kt \) represents a line on waterfall plot, \( b \) adjusts signal bandwidth and \( N(f,t) \) is Gaussian random field with zero mean and identity covariance matrix. Radon transform recovers the signal as a bright spot with a characteristic hourglass profile. For the plots of \( n \times n \) pixel resolution Radon integration amplifies the signal by a factor \( \sim n \) and noise by a factor \( \sim \sqrt{n} \), resulting to typical increase of SNR by a factor \( \sim \sqrt{n} \) relative to the original signal.

Methods of the analysis

In this paper we will apply Radon transform for the analysis of waterfall plots available at setilive.org. The general scheme of the analysis is shown on Fig.2. The main component, Radon transform is defined by the formula:

\[ r(x,a) = \int dy \, s(x \cos a - y \sin a, x \sin a + y \cos a) \]

In numerical estimation the integral has been replaced by a sum over corresponding pixels restricted to the bounds of the waterfall plot. The resulting distribution has been normalized by subtracting the average and dividing to rms. The normalization maps \( \text{rms} \rightarrow 1 \), so that the resulting 2D plot directly represents Radon transform in SNR units.

![Fig.2. Scheme of analysis.](image)

Prescaling step is needed to remap the original waterfall plots to the size appropriate for Radon transform. The original plots are 8-bit grayscale PNG images, coming in two resolutions: 768x384 and 758x410. Since the above described version of Radon transform is suitable for square images, the original plots have been rescaled to 256x256 resolution, using ImageMagick 'convert' tool. In this paper we use direct \( O(n^3) \) algorithm for Radon transform. As an improvement, one can replace it with fast \( O(n^2 \log n) \) Radon transform and its versions [4-6]. They are designed for the images, whose dimensions are either integer powers of two or prime numbers. The prescaling step will be
also necessary for fast transform. Signal selection is the most important step in the chain. The input waterfall plots come in groups, corresponding to simultaneous signal observations in several directions on celestial sphere. Only the signals highly localized on the sphere are interesting, i.e. the signals present in one observation from the group and absent in others, so called 'single beams'. As the first step, only the groups containing 3 observations have been selected, reducing the total number of groups from 6.73e5 to 2.53e5. Then the signals with small Doppler drift were eliminated. Such signals are produced by the sources which don't move (or have a constant radial speed) relative to the receiver and most probably correspond to terrestrial radio sources or geostationary satellites. These signals are represented as vertical lines on waterfall plots, i.e. as nearly zero angles on Radon plots and can be easily eliminated.

![Fig.3. On the left: selection of narrow signal on Radon plot. On the right: vertical correlation on original waterfall plot.](image)

Further, the signals were selected on Radon plots using a method shown on Fig.3 left. Two rectangular contours around a given point were defined. The point is selected if a logical AND of the following conditions took place:

- in the given beam in the given point \( \text{SNR}_0 > 4 \)
- in the given beam everywhere between two contours \( \text{SNR} < \frac{\text{SNR}_0}{2} \)
- in the other two beams everywhere inside external contour \( \text{SNR} < 4 \)

This allows to select narrow signals appearing in a single beam. On Fig.4 left the signals are presented using as coordinates SNR in main beam and maximum of SNR in two other beams. In more detail, in the main beam \( \text{SNR}_0 \) in the given point is taken, while in two other beams the maximum SNR inside the external contour around the given point is taken. The grid step on this plot equals 1. The diagram can be separated to random background, a region of radio frequency interference and a region for potential extraterrestrial signals. The latter region corresponds to strong signal in main beam and weak signals in the other beams. Fig.4 right shows a cumulative distribution plot, where the vertical axis presents the number of selected events possessing SNR in the main beam greater than a value given on the horizontal axis. In appropriate normalization this plot presents an empirical distribution function, a finite sample approximation to cumulative distribution function (CDF), characterizing the probability \( P(\text{SNR} > \text{val}) \) in its tail region.
Fig.4. On the left: separation of RFI and potential ET signals. On the right: distribution of selected signals in comparison with Monte Carlo simulation.

Monte Carlo simulation

For comparison, we show on Fig.4 the results of Monte Carlo simulation, where Radon transform and the same selection criteria were applied to purely random field. Significant difference of these distributions starts at about SNR~6, which can be used as a threshold separating signals from the background.

In more detail, Radon integral contains lines with different number of pixels e.g. comparing locations in central region and corners of waterfall plots. To compensate this effect, the result of Radon transform before global normalization has been divided to sqrt(k), where k is the actual number of pixels counted along the integration line. For Gaussian random field with uncorrelated pixels this will correct SNR for integration lines of different lengths. However, detailed look into original waterfall plots Fig.3 right shows a presence of vertical correlation. Namely, p consequent pixels in vertical direction share the same value, where p varies between 2-4 and most often equals 3. The reason for this correlation could be in the details of measurement procedure, data processing or final representation of images. We will not speculate, but introduce a correction for this effect. For this purpose we assume the groups of p=3 input vertical pixels strongly correlated, count the number of uncorrelated pixels k′ along integration lines and multiply the result of Radon transform to sqrt(k′/k) before global normalization. The same pattern of p=3 vertically correlated random pixels has been also used in MC simulation.

We note also that vertical correlation of Radon plots is not directly related to vertical correlation of waterfall plots. E.g. taking Gaussian random field with identity covariance matrix on input

$$\text{cov}(s(x,y),s(x',y')) = \delta(x-x')\delta(y-y')$$

we obtain the following covariance matrix on output of Radon transform:

$$\text{cov}(r(x,a),r(x',a')) = |\sin(a-a')|^{-1} \chi(x,a;x',a')$$

Here $\chi$ equals 1 if two straight lines defined by $(x,a)$ and $(x',a')$ have intersection inside the domain of random field definition and 0 otherwise. For the random field defined on a square $[-1,1]^2$.
\[ \chi = \theta(|\sin(a-a')|-|x \sin a' - x' \sin a|) \theta(|\sin(a-a')|-|x \cos a' - x' \cos a|) \]

where \( \theta \) is Heaviside step function. In particular, we have

- \( \text{cov}(r(x,a),r(x',a'))=0 \) at \( a=a', \ x\neq x' \), in horizontal direction of Radon plot;
- \( \text{cov}(r(x,a),r(x',a'))\sim |a-a'|^{-1} \) at \( a\rightarrow a', \ x=x', \ |x|<1 \), in vertical direction.

Simulation of random fields with a given covariance matrix can be performed with the methods [7] using principal component analysis (PCA). There are also methods for simulation of random fields with unknown covariance matrix based on numerically efficient singular value decomposition (SVD) of a representative set of samples from the field [8]. These methods can be used as an improvement of MC technique used here for estimation.

In our analysis the processing of real data and MC simulation used the same number of waterfall plots, they were performed on parallel CPUs and were accomplished in the same time. Although such synchronization is not obligatory, since the simulated cumulative distribution could be scaled to the size of real data with an appropriate factor. The results of MC simulation with SNR~6 have a form of bright spots on Radon plots and faint lines on original waterfall plots. They are visually indistinguishable from weak signals we are looking for. The only difference is that for MC results we know definitely that they have random origin. Their small probability was amplified by a large statistics and a narrow filter tuned to the signals of such shape. Although such weak signals cannot be distinguished from random events by their shape, the number of observed signals at SNR~6 should be considered as statistically significant, since it prevails over the estimated number of signals randomly adopting the shape of straight lines.

Thus we leave in selection all signals with Radon's SNR>6. This threshold separates ~1500 events. Our selection algorithm can trigger several events per one signal, because several neighbor pixels of Radon plot can pass selection criteria. In the list below the duplicates will be eliminated, all events originating from the same observation will be represented by only one observation code, comprising 1072 signals in total.

**The results**

The figures below show the brightest signals from the selected set, while the Appendix contains all selected signals. We see that the algorithm works stable also in the presence of several straight lines on the waterfall plot, since they become separable on Radon plot. The lines on the waterfall plots can be reconstructed from selected pixels on Radon plots, actually applying inverse Radon transform to them. As a further improvement of the algorithm, one can apply the inverse transform not to single pixels but to a group of pixels near selected signal on Radon plot or perform non-linear transforms like contrast corrections \( r\rightarrow r' \) on Radon plot and apply the inverse Radon transform to the result.

The signals in Appendix are selected 'with a margin' and can require further filtration:

- Many selected signals possess only a slight slope to the vertical on waterfall plots. Their selection depends on the threshold for vertical lines elimination. Currently this threshold has been selected to \pm 2\ pixels in vertical direction from the center of 256x256 Radon plot, corresponding to 2° slope to the vertical on original waterfall plots.
- Many selected signals have a periodical structure combined in parallel lines, e.g. ed9f on Fig.7. This implies the same source for several related frequencies. The signals of this kind can be attributed to terrestrial or satellite RFI. On Radon plane they create a sequence of
spots located on the same horizontal line. Currently all such signals are preserved in the list. Further the algorithm can be upgraded for automatic recognition of such sequences and their identification as multiple components of the same signal.

- The signals with SNR<7 are almost impossible to recognize on waterfall plots. Their presence is important nevertheless. Firstly, one can see them on Radon plots and this gives enough evidence for their existence. Secondly, the number of reconstructed signals at this level in real data prevails over MC results. These signals should be subjected to further analysis to identify the reason of such excess.

Finally, selected single beams can be subjected to more sophisticated analysis [1], involving the tests for reproducibility of the signal in repeating observations in a given beam, the absence of signal with similar parameters in other beams, not only for simultaneous observations but also in all observations within a certain time frame, etc.

**Conclusion**

We have performed statistical analysis of data stored in SETILive archive, using Radon transform and specially constructed filter for selection of single beams. 28 strong signals and 1072 signals in total were selected at the level SNR>6 in the main beam on Radon plot. This level corresponds to the point where CDF of the signals changes its behavior and starts to differ from the results of Monte Carlo simulation of Gaussian random field.

We will greatly appreciate continuation of SETILive project, which will also make meaningful further developments of the described methods including

- real-time data analysis based on fast Radon transform
- fast inverse Radon transform for reconstruction of selected signals
- parallelization to several processors or/and GPUs
- Monte Carlo simulation of random fields using PCA/SVD
- clustering techniques for detection of multicomponent signals
- distribution of selected signals over observation targets

**References**

[1] Andrew P. V. Siemion et al, A 1.1 to 1.9 GHz SETI Survey of the Kepler Field: I. A Search for Narrow-band Emission from Select Targets, The Astrophysical Journal, 767:94 (13pp), 2013, <arxiv.org/abs/1302.0845>

[2] Stanley R. Deans, SETI, Radon Transforms, and Optimization, Maximum-Entropy and Bayesian Methods in Science and Engineering, Fundamental Theories of Physics, Volume 31-32, 1988, pp 1-17.

[3] Dave Robinson, Radon transform provides alternative waterfall plots, <setiquest.org/blog/radon-transform-provides-alternative-waterfall-plots>

[4] Gregory Beylkin, Discrete Radon Transfrom, IEEE Transactions on Acoustics, Speech, and Signal Processing, Volume ASSP-35, 1987, pp 162-172.

[5] Shekhar Chandra et al, Fast Mojette Transform for Discrete Tomography, Preprint 2010, <arxiv.org/abs/1006.1965>

[6] Finite Transform Library (FTL), <finitetransform.sourceforge.net>

[7] Veit Bayer and Dirk Roos, Non-parametric structural reliability analysis using random fields and robustness evaluation, in WOST 3.0, Weimar, 2006.

[8] Igor Nikitin, Lialia Nikitina, Tanja Clees, Stochastic analysis and nonlinear metamodeling of crash test simulations and their application in automotive design, in Browning, J.E., Ed., Computational engineering: Design, development, and applications, New York, Nova Science Publishers, 2012, pp. 51-74.
Fig. 5. Selected signals with highest SNR. The plots for all 3 beams are shown. Radon plot is shown on the right of every waterfall plot.
Fig. 6. Selected signals with highest SNR (cont'd). Reconstructed signals on waterfall plots are shown.
Fig. 7. Selected signals with highest SNR (cont'd).
