Developments below the Survey Threshold

Mario Santos
for
Jonathan Zwart, Song Chen, Eliab Malefahlo, Russ Taylor &
Matt Jarvis
mariogrs@gmail.com

5 November, 2016

www.uwcastro.org

Plan

- Why bother?
- Ways to beat the survey threshold
- Bayesian approach to modelling the data
- Activity report 2015/2016
- The threat of confusion
- Some caveats

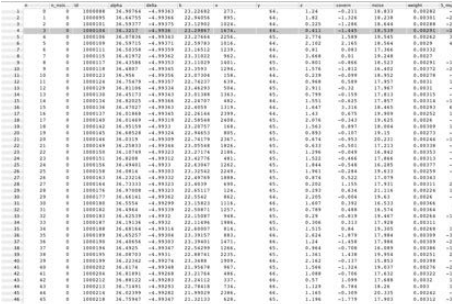
What's on offer?

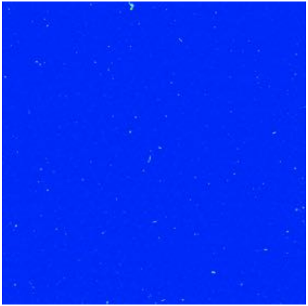
- Counts → effects of confusion on arrays + surveys (cf. 4C)
- For example, MIGHTEE will be **confusion-limited** (return to later)
- We are also seeing confusion effects in e.g. MWA/GLEAM — eventually SKA1-LOW....
- But there's a tonne of science to be done with these data, especially when we have ancilliary (optical/nir) data in hand
- Typically, photo-z's and stellar masses →
  - Luminosity functions
  - (Specific) Star formation rates
  - Far-infrared–radio correlation
  - Spectral indices as function of flux
  - Polarization fraction as function of flux
  - Two-point correlation function
  - HI mass function + evolution...
  - Split by population (using best-fit templates)

Options for beating the noise:

1. **Observe** for as long as $\sqrt{t}$ holds (but **systematics**)
2. **Build** a new telescope
3. **Dedicated** $P(D)$ (Scheuer 1957): **Blind** analysis of map's flux 'deflections' to measure **total** noise contribution from faint confusing sources, and hence count these as function of flux (> 1 source/beam $\rightarrow$ confusion-limited)
4. **Stacking** (various): **Select** target population and use their prior catalogue **positions** to measure fluxes in some other map, even if **undetected** in that map ('many' beams/source please)

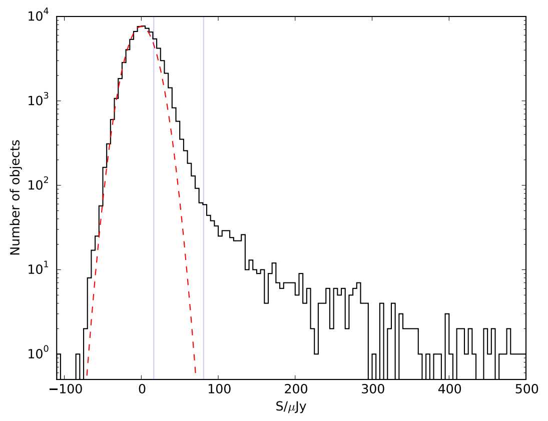
Stacking = Covariance of a map with a catalogue





- Or use ALL the pixels in the map
- ($P(D)$ experiment = arrange for confused data)

Data: Bin → detected, intermediate, noisy



Directly from counts to parameter space

- ‘Beyond stacking’ (Mitchell-Wynne+ 2014)
- ‘Bayesian likelihood analysis’ (Vernstrom+ 2014)
- ‘Parametric stacking’ (Roseboom & Best 2014)
- ‘Far beyond stacking’ (Zwart, Santos & Jarvis 2015)
- …

- Need a generative, parametric model for observed pixel-count distributions/deflections — use Bayes’ theorem for this, with Poisson likelihood function:

Short interlude: Bayes' theorem

$$\boxed{\mathcal{P}\left(\mathbf{\Theta}|\mathbf{D}\right)} = \frac{\mathcal{L}\left(\mathbf{D}|\mathbf{\Theta}\right)\Pi\left(\mathbf{\Theta}\right)}{\boxed{\mathcal{Z}\left(\mathbf{D}\right)}}$$

Use it for:

1. **Parameter estimation** — probably familiar to you (beware: assumes **know** model)
2. **Model selection** (even these **inescapably** need own priors — often neglected...), i.e. Occam's razor

NOT maximum likelihood, etc. — rather, use all available information, explore full posterior distribution via MCMC/nested sampling, propagate uncertainties and correlations 👍

The stacking algorithm for source counts

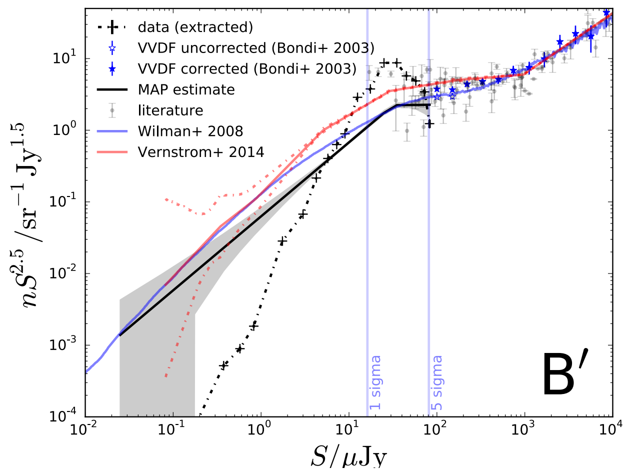
1. start with a radio map
2. use a catalogue of source positions (usually from a survey at other wavelengths) - this is our "target population"
3. extract flux from radio map at source positions (assume only the source flux is contributing + thermal noise)
4. Histogram: bin number of sources per extracted flux bin - our flux bins will have negative values (because of noise and average subtraction from the interferometer). There can also be bins above the typical flux cut (for strong sources) - usually neglect these ones in the fitting.
5. Choose the dN/dS model - it will have to include a minimum flux cut (below which you assume your population does not have any sources) and a maximum flux cut (only needs to be above your maximum data flux cut so to be able to describe the data you observe (noise might "move" high fluxes into measured lower fluxes).
6. fit the model to the obtained histogram using the algorithm and an assumed probability distribution for your histogram values (close to Gaussian for lots of sources in each bin)

Models — whatever you want

- For source counts, use piecewise power laws or polynomials or independent bins/flat bands
- For other quantities, zone data into separate histograms by noise, redshift, polarization, stellar mass etc. and just add log-likelihoods
- Then use Schechters or double PLs or...
- Eventually analyse on a per-object basis (Ocran; hard — and still assumes a perfect map!)

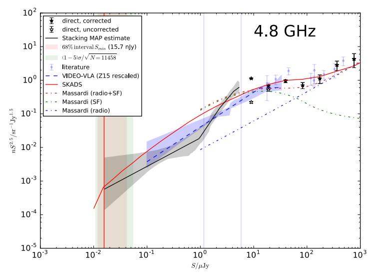
Results

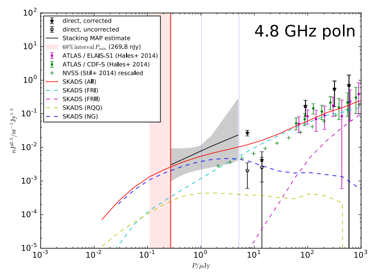
1. VIDEO-VLA 1.4-GHz source counts (Zwart+)



2. ELAIS-N1-JVLA 4.8-GHz *I*/*P* source counts (Zwart+)

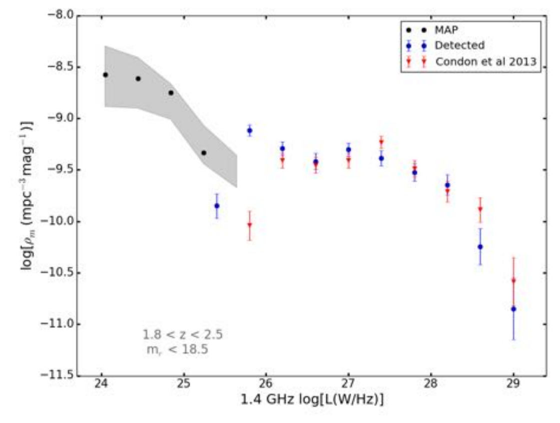
- Added Rice/Rayleigh noise model
- Added multiple noise zones
  → deepest GHz counts to date
  → existence of polarized signal





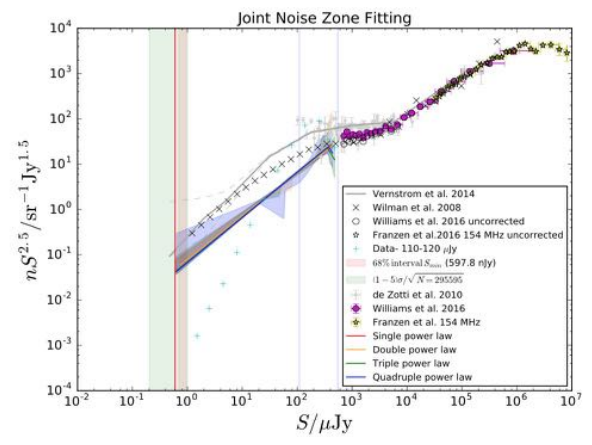
3. SDSS Quasars stacked in FIRST (Malefahlo+)

- Adding luminosity-function generative model



4. The deepest 150-MHz counts (Hale+)

- μJy LOFAR counts!



Systematics

Salient equation (Mitchell-Wynne)

- In order to generate model counts per bin (for sampling), convolve source-count model (however generated) with noise:

$$I_{\mathrm{bin}} = \int_{S_{min}}^{S_{max}} \mathrm{d}S \frac{\mathrm{d}N(S)}{\mathrm{d}S} \int_{S_{m_i}}^{S_{m_i}+\Delta S_{m_i}} \mathrm{d}S_m \frac{1}{\sigma_n\sqrt{2\pi}} \mathrm{e}^{-\frac{(S-S_m)^2}{2\sigma_n^2}}$$

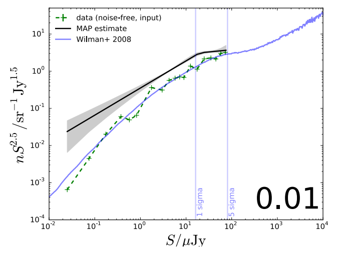
- The noise (red) can be Gaussian in each polarization, or we can incorporate confusion:

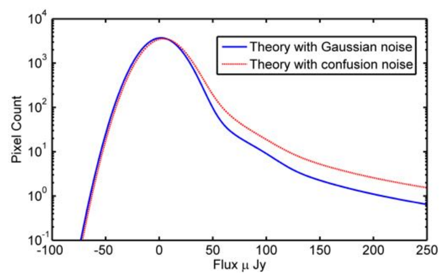
$$I_{\mathrm{bin}} = \int_{S_{min}}^{S_{max}} \mathrm{d}S \frac{\mathrm{d}N(S)}{\mathrm{d}S} \int_{S_{m_i}}^{S_{m_i}+\Delta S_{m_i}} \mathrm{d}S_m P(D \equiv S - S_m)$$

Thinking about confusion now:

- Scheuer derived $P(D)$, idea later developed by Condon
- Analytic expression for counts → confusion noise/histogram (capitalized on by Vernstrom+)
- Fourier analysis → computationally intensive
- So for time being, can just use measured $P(D)$ for OFF pixels
- This will let us pierce the confusion wall with stacking (which goes as $\sqrt{N}$)
- So MIGHTEE is fine
- Eventually fit confusion counts too

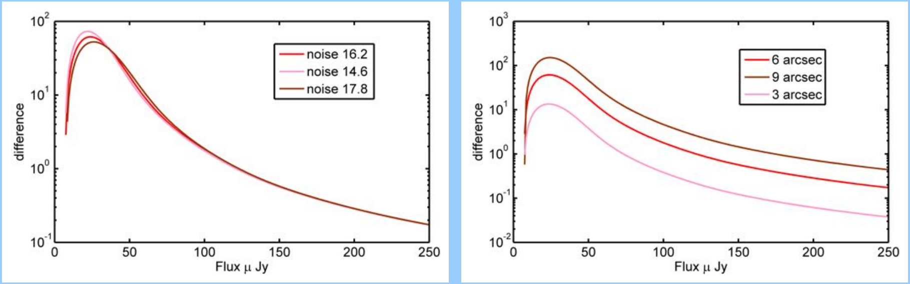
5. Stacking in presence of confusion (Chen+ in prep.)





- Left: Bias noticed by Zwart+ 2015 when stacking SKADS sources assuming Gaussian noise (VIDEO-VLA setup)
- Right: Now correct for effect of confusion

Effects on confusion of thermal noise and resolution



Elephant trap: Caveats & biases (astro-ph.GA/1412.5743)

- **Sky-side**
  - Source clustering and sample variance (Heywood+ 2013)
  - Extended sources... (Vernstrom+ 2015)
  - Parent population (prior-based, so NB selection effects)
  - Confusion
- **Instrument-side**
  - Biases are usually downwards — smearing
  - Resolution bias (oversampling: 20 per cent → 1 per cent)
  - CLEAN/snapshot bias ($< 1\mu$Jy — Bondi+ 2003)
  - Astrometry (X-band)
  - The eggbox — varying map noise
  - Noise model
  - Calibration, sidelobe confusion, ghosts etc. (Smirnov+)
  - Use of a single-point statistic

Prospects are good

- Santos says: Don't detect....
- Stacking gets us orders of magnitude deeper
- The days of median stacking are behind us
- Bayesian framework for sub-threshold work now well established
- Code available — BAYESTACK (`jz@uwcastro.org`)
- Not short of applications
- Full end-to-end imaging simulations are (still) the Achilles' heel — but this is generally true
- Ultimately only a visibility analysis (BIRO/RIME) will hand systematics correctly