Telescopes don’t make catalogues!

Slide title stolen from D.W. Hogg and D. Lang, EAS Publication Series 45, 351 (2011)
Image: Sloan Digital Sky Survey, DR12
People make catalogues

<table>
<thead>
<tr>
<th>RA (J2000)</th>
<th>DEC (J2000)</th>
<th>g</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>229.4351</td>
<td>2.010923</td>
<td>19.344</td>
<td>19.114</td>
</tr>
<tr>
<td>229.4353</td>
<td>1.990166</td>
<td>23.07</td>
<td>21.894</td>
</tr>
<tr>
<td>229.4358</td>
<td>2.033374</td>
<td>21.809</td>
<td>21.252</td>
</tr>
<tr>
<td>229.4361</td>
<td>2.070269</td>
<td>20.107</td>
<td>19.737</td>
</tr>
<tr>
<td>229.4362</td>
<td>1.997957</td>
<td>22.894</td>
<td>21.365</td>
</tr>
<tr>
<td>229.4364</td>
<td>2.048578</td>
<td>22.386</td>
<td>21.656</td>
</tr>
<tr>
<td>229.4366</td>
<td>2.053515</td>
<td>20.853</td>
<td>21.161</td>
</tr>
<tr>
<td>229.4369</td>
<td>2.105316</td>
<td>21.827</td>
<td>21.161</td>
</tr>
<tr>
<td>229.4369</td>
<td>2.043476</td>
<td>23.067</td>
<td>21.656</td>
</tr>
<tr>
<td>229.437</td>
<td>2.051732</td>
<td>19.96</td>
<td>19.737</td>
</tr>
<tr>
<td>229.4371</td>
<td>2.102266</td>
<td>20.813</td>
<td>21.161</td>
</tr>
<tr>
<td>229.4373</td>
<td>2.052342</td>
<td>20.785</td>
<td>21.656</td>
</tr>
<tr>
<td>229.4374</td>
<td>1.996688</td>
<td>21.161</td>
<td>21.656</td>
</tr>
<tr>
<td>229.4376</td>
<td>2.13321</td>
<td>22.476</td>
<td>21.161</td>
</tr>
<tr>
<td>229.4378</td>
<td>2.039289</td>
<td>20.883</td>
<td>21.656</td>
</tr>
<tr>
<td>229.438</td>
<td>2.077996</td>
<td>22.682</td>
<td>21.656</td>
</tr>
<tr>
<td>229.438</td>
<td>2.043483</td>
<td>22.884</td>
<td>21.656</td>
</tr>
<tr>
<td>229.4381</td>
<td>2.045585</td>
<td>20.111</td>
<td>19.737</td>
</tr>
<tr>
<td>229.4382</td>
<td>2.011463</td>
<td>22.069</td>
<td>21.656</td>
</tr>
<tr>
<td>229.4382</td>
<td>2.029807</td>
<td>19.625</td>
<td>19.737</td>
</tr>
<tr>
<td>229.4382</td>
<td>2.030182</td>
<td>17.835</td>
<td>17.245</td>
</tr>
<tr>
<td>229.4385</td>
<td>2.157053</td>
<td>22.193</td>
<td>21.877</td>
</tr>
<tr>
<td>229.4385</td>
<td>2.147021</td>
<td>22.492</td>
<td>21.546</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
(Deterministic) Catalogues

- A (deterministic) catalogue is a list of point source candidates above some inclusion threshold $TS_{incl}$

$$\text{Data, } TS_{incl} \rightarrow \{\ell_i \pm \sigma_{\ell_i}, \delta_i \pm \sigma_{\delta_i}, F_i \pm \sigma_{F_i}\}_{i=1}^N$$

- **Inclusion threshold = detection threshold:**
  Almost all catalogue sources are true sources
  But faint true sources are not in the catalogue

- **Inclusion threshold < detection threshold:**
  More faint true sources are included in the catalogue
  But many catalogue sources are not true sources
  The data is overfitted

- Overlapping point sources may not be deblended
Probabilistic Catalogues

• A probabilistic catalogue is a posterior probability distribution over the space of lists of point source candidates

\[ P\left(\{\ell_i, \vartheta_i, F_i\}_{i=1}^N \mid \text{Data}\right) = \pi\left(\{\ell_i, \vartheta_i, F_i\}_{i=1}^N\right) \mathcal{L}(\text{Data} \mid \{\ell_i, \vartheta_i, F_i\}_{i=1}^N) \]

• Sampling the probabilistic catalogue provides an ensemble of catalogues inferred from the data

Why Probabilistic Catalogues?

• The reality of a single faint point source candidate will be very uncertain, but the properties of a faint population are constrained

• The uncertainty in deblending sources with overlapping PSFs can be captured

• Provides a framework to marginalize over uncertainties (modelling, instrumental, calibration, etc.)

• Probabilistic cataloguing more fully captures the information contained in the data and the inherent degeneracies of point source identification
Application I: *Fermi* High Latitude

- North Galactic Pole 20° × 20° \(N_{pix} = 29\,880\)
- 3 energy bins: 0.3-1 GeV, 1-3 GeV, 3-10 GeV
- Region includes 108 3FGL sources
- Run with ~250 CPU-hours
- Diffuse sources:
  - Galactic diffuse emission
  - Isotropic emission
- Point source population:
  - Mostly distant active galaxies
  - Assumed to be isotropically distributed
  - Unknown flux distribution parameterized as power law
Catalogue Samples
Flux and Colour Distributions

![Graphs showing flux and colour distributions.](image-url)
Application II: SDSS Globular Cluster

- Messier 2 40" × 40" ($N_{\text{pix}} = 10\,000$)
- Region includes 337 DAOPhot sources
- Run with ~250 CPU-hours
- Region has also been observed with HST, which has better angular resolution, identifying 1,000 sources
Deterministic Catalogue of SDSS Data

Deterministic Catalogue of HST Data
Probabilistic Catalogue of SDSS Data

SDSS DAOPhot
HST DAOPhot
SDSS PCat-Dnest
Completeness
Reversible Jump MCMC

• Allows proposals to change dimensionality of model
  • Move $m$ takes $x$ and generates auxillary $u$ to propose $x'$
  • Move $m'$ takes $x'$ and generates auxillary $u'$ to propose $x$
  • $\dim x + \dim u = \dim x' + \dim u'$ and $(x,u) \leftrightarrow (x',u')$ one-to-one

$$\alpha(x \rightarrow x') = \min \left(1, \frac{\pi(x') \mathcal{L}(x'|D) j_m(x') g(u')}{\pi(x) \mathcal{L}(x|D) j_m(x) g(u)} \left| \frac{\partial(x',u')}{\partial(x,u)} \right| \right)$$

• For example, birth/death between $x = \{x_1, ..., x_N\}$ and $x' = \{x_1, ..., x_{N+1}\}$ has $u = x_{N+1}$ and $u' = \emptyset$
  • If birth and death equally likely, sources independent in prior and new source $x_2$ generated from prior

$$\alpha(x \rightarrow x') = \min \left(1, \frac{\pi(N+1) \mathcal{L}(x'|D)}{\pi(N) \mathcal{L}(x|D)} \right)$$

P.J. Green, Biometrika 82, 711 (1995)
Catalogue Priors

- Prior that sources are independent and described by population parameters $\beta$:

$$\pi(\{\ell_i, \kappa_i, F_i\}_{i=1}^N, \beta) = \pi(\beta)\pi(N|\beta) \prod_{i=1}^N \pi(\ell_i, \kappa_i, F_i|\beta)$$

- $\beta$ can describe both spatial and flux distributions

- What should the prior on the number of sources look like? What do we mean by “the number of sources”?

  *How many sources are there with a flux above $F_{\text{min}}$?*

- Prior on $N$ through putting a log uniform prior on expected number of sources $\langle N \rangle$?

  $$\log \frac{\pi(N + 1)}{\pi(N)} = \log N - \log(N + 1) \approx -\frac{1}{N}$$
Source Number Prior

• But is this prior enough to counteract the fact that models with more sources will fit better?

• What about a prior that penalizes the $(N + 1)^{\text{th}}$ source based on the expected improvement in $\chi^2$ under the null hypothesis that there are $N$ sources?

\[
\log \frac{\pi(N + 1)}{\pi(N)} = -\frac{3}{2}
\]

*How many sources meaningfully affect the current data?*

*What is the most compact representation of the data?*
Conclusion

- Probabilistic catalogue samples are an ensemble of catalogues inferred from the data

- A point source population can be distinguished from a diffuse source, even if the individual sources are below the detection threshold

- Overlapping point sources can be better deblended

- This ensemble of catalogues captures the inherent degeneracies of point source identification