Mon. Not. R. Astron. Soc. **000**, 1–?? (2013) Printed 15 August 2013 (MN LATEX style file v2.2)

Milky Way Project : Clouds and Holes Paper

R. J. Simpson[1⋆], C. J. Lintott[1,2], C. E. North[3], and friends...

[1] *Oxford Astrophysics, Denys Wilkinson Building, Keble Road, Oxford, OX1 3RH, UK*
[2] *Astronomy Department, Adler Planetarium, 1300 S. Lake Shore Drive, Chicago, IL 60605, USA*
[3] *Department of Physics and Astronomy, Cardiff University, 5 The Parade, Cardiff, CF24 3YB*

Accepted 20xx. Received 20xx.

ABSTRACT
Abstract

1 INTRODUCTION

Infrared dark clouds (IRDCs) were first observed as dark regions silhouetted against the mid-infrared (MIR) background (**?**) by *Infrared Space Observatory* (**?**) and *Midcourse Space Experiment* (**?**). Subsequent observations showed them to have low temperatures and high densities ($T \lesssim$K, $n_H > 10^5\,\mathrm{cm}^{-}3$, e.g. **???**). The IRDC absorbs the background light and causes a dip in the MIR sky brightness. They are thought to be the earliest observable formation stages high-mass stars and stellar clusters.

Following the initial surveys, the MIPSGAL and GLIMPSE surveys used the IRAC and MIPS instuments on *Spitzer*allowed the compilation of larger and more complete catalogues (e.g. **?**). With MIR data alone, however, it is impossible to distinguish this absorption from a region of inherently lower background emission. Far-Infared (FIR) observations allow IRDCs, which appear bright at wavelengths above 24 µm, to be distinguished from regions of lower emission, which remain dark – a 'hole in the sky' (e.g. **?**). It is important to identifying which of the candidate IRDCs in the catalogues are 'true' dark clouds, and which are simply holes in the sky. Such interlopers would bias any statistical studies of IRDCs based on these catalogues.

The HIGAL survey, using the SPIRE and PACS instrument on-board the *Herschel Space Observatory*, covered a $|b| \leqslant 1$ deg strip of the entire Galactic Plane at 70, 160, 250, 350 and 500 µm. Using the 8 and 24 µm data from GLIMPSE and MIPSGAL, combined with the longer wavelength data from HIGAL, it is, in princple, possible to distinguish which objects are clouds and which are holes.

1.1 Input catalogue

This study uses candidate IRDCs in the inner regions of the Galactic Plane (the first and fourth quadrants). We use two catalogues based on MIR data from the MSX and *Spitzer* surveys (**??**), and add to that FIR data from the *Herschel*HI-GAL survey. The **?** catalogue (henceforth S06) uses 8 µm MSX data to identify over 10000 IRDCs in the region defined by $|l| < 90°$ and $|b| < 5°$. This was restricted to IRDCs greater than 36" in size (limited by the 20" resolution of MSX) and with a sensitivity of 1.2 MJy/sr. The **?** catalogue (henceforth PF09), used *Spitzer* 8 and 24 µm data in the region defined by $10° < |l| < 65°$ and $|b| < 1°$. While *Spitzer* data had better resolution and sensitivity (2" and 0.3 MJy/sr at 8 µm), it covered a smaller region of the Galactic Plane. The Hi-GAL survey covers the full Galactic Plane within $|b| < 1°$, with a resolution of $\sim 18''$ at 250 µm.

Sources included in this study require coverage at both 8 and 24 µm. They are selected from PF09, along with sources from S06 which are not within 25" of a PF09 source. This results in 9931 sources from PF09 (4123 in the range $0 < l < 90$ and 5808 in the range $270 < l < 360$) and 5731 from S06 (2672 in the range $0 < l < 90$ and 3059 in the range $270 < l < 360$).

2 ANALYSIS

Brief outline of analysis

2.1 Milky Way Project

The Milky Way Project[1] was established in 2010 as a citizen science interface to data from the *Spitzer* GLIMPSE survey primarily as a search for 'bubbles' associated with massive star formation. This effort was successful, and a catalogue of more than 5000 such bubbles which expanded on previous efforts by professional astronomers was published by **?** and used for a statistical analysis of bubble distribution by **?**. Inspired by this success, a second interface was added to the site in order to address the problem of identifying true IRDCs.

As with the previous interface, this new part of the site[2] makes use of the Zooniverse Application Programming Interface (API) originally built for Galaxy Zoo (**?**) and which supports a large number of similar citizen science projects. This API is primarily responsible for serving images and recording classifications provided by volunteers, who are required to be logged in for their work to be recorded. The

[1] http://www.milkywayproject.org
[2] http://www.milkywayproject.org/clouds

Figure 1. The interface for the Clouds part of the Milky Way Project, as seen by classifiers who had to sort each image into one of three categories.

interface itself is built in JavaScript and HTML5. Following a short tutorial, an image is selected from the database[3] and presented to the volunteer who may label it as a CLOUD, a HOLE or an INTERMEDIATE case by selecting one of three buttons, as shown in figure **??**. Examples of both clouds and holes are provided during the tutorial phase and can be reviewed at any time; these examples are shown in **??** in both colour schemes available to classifiers. Once an image is classified, the volunteer is shown another image and presented with the opportunity to discuss the first image with other classifers[4].

DETAILS OF IMAGES AND IMAGE CONSTRUCTION.. Participants can chose to alter the colour palette presented (DO WE RECORD IF PEOPLE USE THIS OPTION?)

The clouds project was launched on XXXXXXX and ran until XXXXXXX, collecting 1.1 million classifications. 3544 logged-in users provided classifications. However, approximately half of the classifications were received from those who were not logged in. The most active user has completed 59020 classifications and is one of three users to have seen every image provided. (HOW DO WE HANDLE THE REPEATS IN DATA REDUCTION). 3253 classifiers provided more than five classifications (91.9%, compared to 25% in the previous incarnation of the Milky Way Project), 1843 more than fifty (52.0% compared to 5.7%) and 168 five hundred classifications or more.

DETAILS OF NON-LOGGED IN USERS

MWP interface and initial results (or in results section?) User and classification numbers and duration of dataset used

Raw classifications? Histogram of results?

Thumnail examples of clouds, holes and unknowns

[3] Volunteers see an image they have not yet classified, selected randomly from those with the fewest classifications in the database. This algorithm for task assignment has the advantage of ensuring that images have approximately the same number of classifications at all times, facilitating preliminary data analysis.
[4] See http://talk.milkywayproject.org

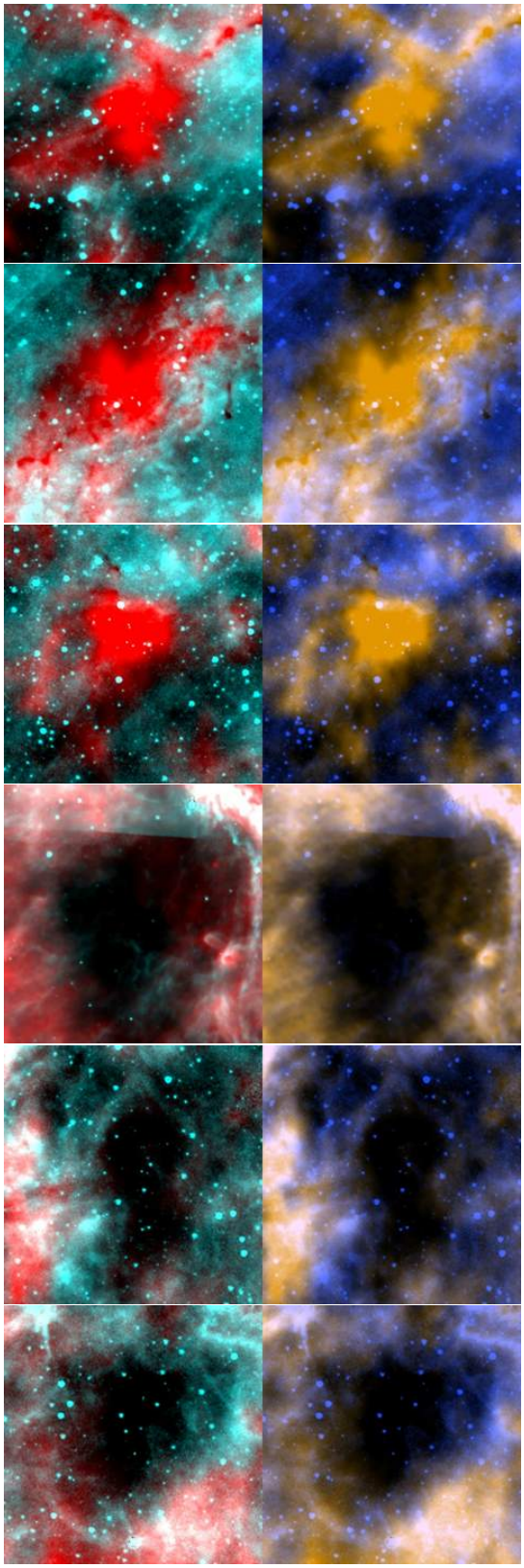

Figure 2. Top three rows : Three of the nine examples of true clouds used in the tutorial. Bottom three rows : Examples of holes given in the tutorial. The tutorial did not include examples of 'intermediate' images.

2.2 Experts and Training data

Definition of training data definition

Thumbnails of training data and classifications

Expert versus general results (plot)

2.3 Analysis sequence

2.3.1 What has been done

All the 1.1 million classifications are run through sequentially in time.

Issues caused because users have three options - cloud, hole or intermediate. (see GZ argument about non-classified galaxies). All clouds start with an initial score of 0.5, which is updated based on the user's skill level. Skill level is determined by users encountering training data.

Note that the tutorial doesn't really say much about intermediates - or at least it only says the button exists. We should therefore expect a wide range of user behaviours - including those who use it as a 'don't know' alongside those who genuinely only use it for intermediate. This means that weighting is important for this category (?).

The training data is currently a set of 'definite' clouds and 'definite' holes assembled by RS - these were assembled by expert review of the top (approx) 200 at each end. NEED MORE RIGOROUS

How do we treat the intermediate classifications?

Option A : A third genuine category - therefore need expert data with all three classifications in order to distinguish between the confused and those who are classifying something as intermediate. Cuts are needed for each population producing three lists (which might overlap).

Option B : Treat cloud to hole as a continuous distribution where somethings are definitely (p=1) clouds, some definitely (p=0) holes but most lie somewhere in between. In this model intermediate votes signify p=0.5. In other words we're forcing a discrete estimate of probability.This produces a catalogue including all the objects with the best value probability for each one to which thresholds can be applied.

(Option C : Ignore the intermediate classifications) We might want to test the effect of adding the intermediate button).

Possible changes to training set :

i. Add a set of 'definite' intermediate objects reviewed by experts as existing set. Give these the 'correct' value of 0.5 and adjust user skill levels accordingly.

ii. Get N experts to classify - award 1 for cloud, 0.5 for intermediate, 0 for hole, take average - adjust user skill level for being close to the average. Close to be defined (might want N to be even so 0.5 can be a 'correct' answer).

iii. Get expert assign numerical value (e.g. 0.7). This won't work as people - even experts - are crap at this (and because it's not really a linear continuous scale).

The procedure Growth charts Results of MC runs Any threshholds

3 RESULTS

Results from analyis (cloudiness chart)

Histogram of classifications before & after

Thumnail examples of clouds, holes and unknowns

Full table of results

4 CONCLUSIONS

5 ACKNOWLEDGEMENTS

This publication has been made possible by the participation of more than 40,000 volunteers on the Milky Way Project. Their contributions are acknowledged individually at http://www.milkywayproject.org/authors. The Milky Way Project, and R.J.S. were supported by The Leverhulme Trust. Development of the MWP was partly supported by the National Science Foundation CDI grant: DRL-0941610.

This work is based on observations made with the Spitzer Space Telescope, which is operated by the Jet Propulsion Laboratory, California Institute of Technology under a contract with NASA. This research has made use of the SIMBAD database, operated at CDS, Strasbourg, France.

Herschel is an ESA space observatory with science instruments provided by European-led Principal Investigator consortia and with important participation from NASA.

SPIRE has been developed by a consortium of institutes led by Cardiff Univ. (UK) and including: Univ. Lethbridge (Canada); NAOC (China); CEA, LAM (France); IFSI, Univ. Padua (Italy); IAC (Spain); Stockholm Observatory (Sweden); Imperial College London, RAL, UCL-MSSL, UKATC, Univ. Sussex (UK); and Caltech, JPL, NHSC, Univ. Colorado (USA). This development has been supported by national funding agencies: CSA (Canada); NAOC (China); CEA, CNES, CNRS (France); ASI (Italy); MCINN (Spain); SNSB (Sweden); STFC, UKSA (UK); and NASA (USA).

This research made use of APLpy, an open-source plotting package for Python hosted at http://aplpy.github.com