Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **446,** 1551–1566 (2015) doi:10.1093/mnras/stu2055

The SAMI Galaxy Survey: cubism and covariance, putting round pegs into square holes

R. Sharp,[1,2★] J. T. Allen,[2,3] L. M. R. Fogarty,[2,3] S. M. Croom,[2,3] L. Cortese,[4] A. W. Green,[5] J. Nielsen,[1] S. N. Richards,[2,3,5] N. Scott,[2,3] E. N. Taylor,[6] L. A. Barnes,[3] A. E. Bauer,[5] M. Birchall,[5] J. Bland-Hawthorn,[2,3] J. V. Bloom,[2,3] S. Brough,[5] J. J. Bryant,[2,3,5] G. N. Cecil,[7] M. Colless,[1] W. J. Couch,[5] M. J. Drinkwater,[8] S. Driver,[9] C. Foster,[5] M. Goodwin,[5] M. L. P. Gunawardhana,[10] I.-T. Ho,[11] E. J. Hampton,[1] A. M. Hopkins,[5] H. Jones,[12] I. S. Konstantopoulos,[2,5] J. S. Lawrence,[5] S. K. Leslie,[1] G. F. Lewis,[3] J. Liske,[13] Á. R. López-Sánchez,[5,14] N. P. F. Lorente,[5] R. McElroy,[2,3] A. M. Medling,[1] S. Mahajan,[8] J. Mould,[4] Q. Parker,[5,14] M. B. Pracy,[3] D. Obreschkow,[9] M. S. Owers,[5] A. L. Schaefer,[2,3,5] S. M. Sweet,[1,8] A. D. Thomas,[8] C. Tonini[6] and C. J. Walcher[15]

[1]*Research School of Astronomy & Astrophysics, Australian National University, Canberra, ACT 2611, Australia*
[2]*Australian Research Council (ARC) Centre of Excellence for All-sky Astrophysics (CAASTRO)*
[3]*Sydney Institute for Astronomy (SIfA), School of Physics, The University of Sydney, NSW 2006, Australia*
[4]*Centre for Astrophysics & Supercomputing, Swinburne University of Technology, Mail H30 PO Box 218, Hawthorn, VIC 3122, Australia*
[5]*The Australian Astronomical Observatory, PO Box 915, North Ryde, NSW 1670, Australia*
[6]*School of Physics, The University of Melbourne, VIC 3010, Australia*
[7]*Department Physics & Astronomy, University of North Carolina, Chapel Hill, NC 27599, USA*
[8]*School of Mathematics and Physics, University of Queensland, Brisbane, QLD 4072, Australia*
[9]*ICRAR M468, UWA, 35 Stirling Highway, Crawley, WA 6009, Australia*
[10]*Institute for Computational Cosmology, Department of Physics, Durham University, South Road, Durham DH1 3LE, UK*
[11]*Institute for Astronomy, University of Hawaii, 2680 Woodlawn Drive, Honolulu, HI 96822, USA*
[12]*School of Physics, Monash University, VIC 3800, Australia*
[13]*European Southern Observatory, Karl-Schwarzschild-Str. 2, D-85748 Garching bei München, Germany*
[14]*Department of Physics and Astronomy, Macquarie University, NSW 2109, Australia*
[15]*Leibniz-Institut für Astrophysik Potsdam (AIP), An der Sternwarte 16, D-14482 Potsdam, Germany*

Accepted 2014 September 30. Received 2014 September 30; in original form 2014 July 18

ABSTRACT

We present a methodology for the regularization and combination of sparse sampled and irregularly gridded observations from fibre-optic multiobject integral field spectroscopy. The approach minimizes interpolation and retains image resolution on combining subpixel dithered data. We discuss the methodology in the context of the Sydney–AAO multiobject integral field spectrograph (SAMI) Galaxy Survey underway at the Anglo-Australian Telescope. The SAMI instrument uses 13 fibre bundles to perform high-multiplex integral field spectroscopy across a 1° diameter field of view. The SAMI Galaxy Survey is targeting ∼3000 galaxies drawn from the full range of galaxy environments. We demonstrate the subcritical sampling of the seeing and incomplete fill factor for the integral field bundles results in only a 10 per cent degradation in the final image resolution recovered. We also implement a new methodology for tracking covariance between elements of the resulting data cubes which retains 90 per cent of the covariance information while incurring only a modest increase in the survey data volume.

Key words: instrumentation: spectrographs – methods: data analysis – techniques: imaging spectroscopy.

★ E-mail: rob.sharp@anu.edu.au

1 INTRODUCTION

For two decades, single-fibre multiobject spectrographs dominated galaxy redshift surveys (Huchra, Vogeley & Geller 1999; York et al. 2000; Colless et al. 2001; Eisenstein et al. 2005; Jones et al. 2009; Drinkwater et al. 2010; Driver et al. 2011). More than two million galaxies now have accurate redshift measurements. These surveys have taught us a great deal about large-scale structure and how the bolometric properties of galaxies evolve with cosmic time. In recent years, broad-band photometric surveys have revealed that there is much to be learnt from the spatially resolved properties of large galaxy samples (Driver et al. 2006; Abazajian et al. 2009) in particular, how these properties vary with the large-scale environment (Welikala et al. 2008; Blanton & Moustakas 2009). What is missing from these surveys is knowledge of the kinematics, ages and metallicity of prominent stellar populations across each galaxy, plus the emission line intensities and kinematics that provide insight on the star formation, dynamics and chemical state of the galaxy.

Integral field and Fabry–Perot spectrographs provide the necessary 3D (2D spatial, 1D spectral) information but these are typically designed as single-target instruments (e.g. HIFI and PMAS; Bland & Tully 1989; Roth et al. 2005) and modest survey samples (Veilleux et al. 2003; Sharp & Bland-Hawthorn 2010; Rich et al. 2012; Brough et al. 2013). Recent integral field surveys have managed to observe of the order of 200–300 galaxies over a long observing campaign (e.g. SAURON; Bacon et al. 2001, ATLAS3D; Cappellari et al. 2011, CALIFA; Sánchez et al. 2012).

With a view to obtaining integral field data on thousands of galaxies, the Sydney–AAO multiobject integral field spectrograph (SAMI) concept was born (Croom et al. 2012). This multiobject instrument uses a new kind of fibre bundle – the hexabundle (Bland-Hawthorn 2011; Bryant et al. 2011) – in order to achieve spatially resolved 3D spectroscopy of up to 13 galaxies at a time over a 1°diameter field. The SAMI Galaxy Survey will observe more than 3000 galaxies in a 3 yr campaign. This instrument is proposed to be extended to 50–100 bundles in its next incarnation with a view to obtaining data on a far larger sample of objects (Lawrence et al. 2012). One challenge of the hexabundle technique is its irregular fibre format of close-packed circular fibres (Fig. 1) which for many applications must be reformatted to give well-formed data for uniform analysis. Indeed the same issues are faced by any imaging system if none integer spaxel shifts are employed in dithering. In this paper, we present the resampling solution adopted, along with other processing steps required to remove the instrumental signature, for the SAMI Galaxy Survey.

The code to perform the data cube generation described in this paper is available from the Astrophysics Source Code Library as project asci:1407.006.[1]

2 THE SYDNEY–AAO MULTIOBJECT INTEGRAL FIELD SPECTROGRAPH

The motivation for the SAMI instrument is described by Croom et al. (2012) with technical specification provided in Bryant (2012) and updated in Bryant et al. (2014b). The first explorations of its scientific capabilities are provided by Fogarty et al. (2012, 2014), Ho et al. (2014), and Richards et al. (2014).

[1] Astrophysics Source Code Library, asci:1407.006, http://ascl.net/1407.006 (Allen et al. 2014)

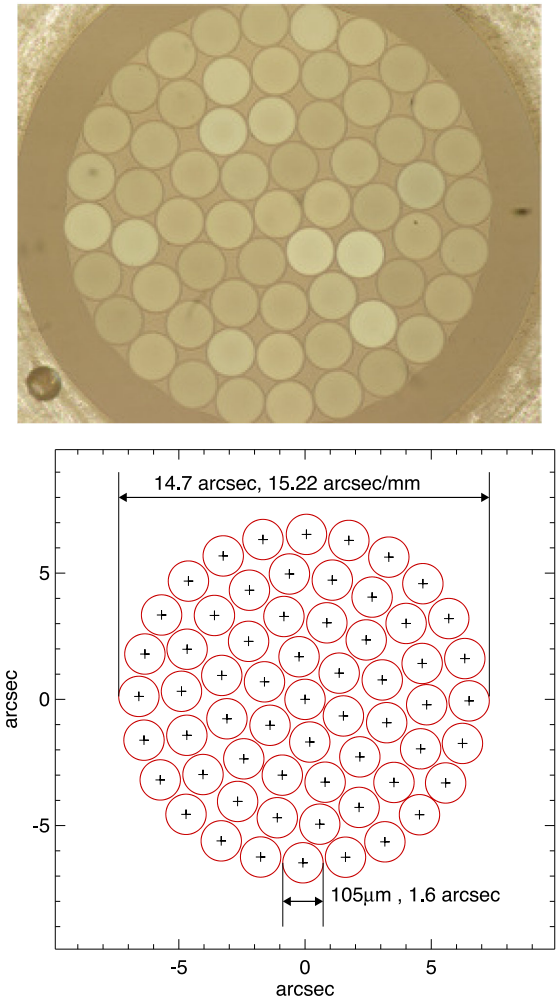

Figure 1. The fibre core arrangement is shown for one of the SAMI fibre IFUs. A quality control image of a bundle, taken during production of a hexabundle, is shown along with a graphical fibre mapping. The none uniform fibre illumination pattern in the image is a product of the back illumination used for the photograph and is not representative of the final bundle transmission. A number of defocused dust artefacts on the laboratory camera are also seen. Each fibre has a core diameter of 105 μm, with the cladding thinned to 110±1 μm over the first 50 mm of each fibre. This allows a tight fibre packing, creating a semiregular IFU array with a filling factor of 73 per cent.

Briefly, the SAMI system uses lightly fused fibre bundles to create self-contained fibre integral field units (IFUs). Each of the 13 SAMI fibre bundles contains a close packed array of 61 optical fibres with individual fibre-core diameters of 1.6 arcsec (Fig. 1). The confinement of the bundle by a circular outer form arranges the fibres into four concentric rings around the central fibre, rather than a hexagonal packing (Bryant et al. 2014a). The fibre bundles populate the 1° diameter focal plane of the $F/3.4$ triplet corrector at the Anglo-Australian Telescope (AAT) via a plug-plate system. Each fibre bundle has a fill factor of 73 per cent over a field of view of ∼15 arcsec diameter. The 13 IFU bundles, each of 61 fibres, and 26 independent blank-sky fibres for sky subtraction, feed the AAOmega spectrograph (Saunders et al., 2004; Sharp et al. 2006). Using the 580V and 1000R gratings, the dual-beam AAOmega spectrograph provides wavelength coverage in two bands, 3700–5700 Å at a spectral resolution of $R \sim 1730$ and 6250–7350 Å at

$R \sim 4500$. The SAMI galaxy survey will observe more than 3000 galaxies in a 3 yr campaign as described in Bryant et al. (2014b).

3 BASIC DATA PROCESSING

A detailed description of the fibre spectroscopy data reduction software 2DFDR, whose usage is common to data taken in all modes of the AAOmega spectrograph, is presented by Hopkins et al. (2013) who provide detail of AAOmega data processing for the GAMA survey programme (Driver et al. 2011). 2DFDR carries out all the reduction steps up to the point of generating Row-Stacked Spectra (RSS) which are wavelength calibrated and sky subtracted with the basic instrumental signatures removed. The RSS frames are 2D images with one row per fibre spectrum (along with the associated variance information and source details in image and binary table extensions) for each observation with the SAMI IFUs. The description of how these are flux calibrated and converted into data cubes is given in Sections 4 and 5 below. Here we will describe the steps required to produce the RSS frames within the 2DFDR package.

The first stage is to subtract bias and dark frames to correct a number of errant CCD pixels. Both arms of the dual-beam AAOmega spectrograph suffer from a number of extended regions of bad columns whose charge transfer inefficiency effects associated with hot pixels can be compensated for by bias/dark correction.[2] An overscan correction is also applied, subtracting the bias level in each frame. After this, each frame is divided by a *detector flat* that is generated by averaging (typically >30) fibre flats for which the spectrograph has been defocused so that the illumination is relatively uniform. These frames are then filtered to remove large-scale variations, leaving only smaller scale pixel-to-pixel flat-field variations. Charge spots due to cosmic rays are removed from each individual science frame using a tuned implementation of the LaCosmic routine (van Dokkum 2001; Husemann et al. 2012). An optimization of the parameters was performed to ensure high confidence in cosmic ray rejection with minimal impact on fibre spectra.

The next stage is to trace the fibre locations across the detector (generating a so-called *tramline map* giving the pixel-by-pixel $[x, y]$ location of each fibre). This is a crucial step as good extraction of 1D spectra from the 2D data frame is contingent on accurately mapping the fibre positions and profiles across the detector. This is performed using a fibre flat-field frame taken using a quartz–halogen lamp that illuminates a white spot on the AAT dome. The fibre intensity is traced in a two stage process. First, the fibre peaks are identified and fitted approximately using a quadratic fit to the 3 pixels around each peak (this gives positions accurate at the ∼0.1 pixel level). Then as a second stage (newly implemented for the SAMI pipeline) we implement an algorithm that assumes a Gaussian fibre profile (a good approximation to SAMI fibres in AAOmega) and fits five Gaussians (the central one and two either side) to precisely determine both the centre and width (to be used later for optimal extraction) of the fibre profile. When fitting we integrate the Gaussian model across each pixel. A robust fourth order polynomial is then fit to both the tramline and width of each fibre as a function of spectral pixel. The multiple Gaussians are required as the close packing of fibres causes the wings of their light profiles to overlap at the ∼5–10 per cent level so that adjacent fibres influence the measured position and centre of the fibre in question (Sharp & Birchall 2010).

Our approach allows two systematic effects present in the fibre traces to be addressed. The first is that the AAOmega CCDs, 2×4k E2V detectors, are manufactured using a lithography mask of 1024×512 pixels. There are errors in the relative positioning of the mask on different parts of the detector, causing discontinuities in the fibre traces between pixel 1024 and 1025 (in the spectral direction). Adding a step when fitting the fibre trace allows us to accurately measure these lithography alignment errors, we find the typical uncertainty on an individual fibre is $\simeq 0.01$ pixels. Because there are ∼100 fibres per lithography block, the mean error per block can be derived at ∼0.001 pixel precision. The worst lithography errors were found to be $\simeq 0.07$ pixels, but most are substantially less than that $\simeq 0.01$–0.02 pixels. A small region of a fibre flat-field is shown in Fig. 2. The residual image (the difference between the input and model flat-field data) without correction for lithography errors (middle panel) shows clear discontinuities that are removed after the fibre traces are corrected for this error (lower panel). In Fig. 3 we plot the lithography errors measured for each fibre and the median values in each block for a typical fibre flat-field from the AAOmega red arm.

A second systematic discovered was a slow shift in position of the fibre traces of approximately 0.02–0.03 pixels per hour. Investigation showed this to be due to a time-varying gravitational torque on the AAOmega cameras caused by the slow boiling off of the liquid nitrogen in the dewers attached to the cameras. Comparison of tram-line maps to the residuals from extracted data frames suggests that our typical total tramline map error is $\simeq 0.02$ pixels, and never worse than 0.05 pixels.

After measuring the tramline map and fibre profile widths, the next stage is extraction of the flux from the 2D image to generate a 1D spectrum for each fibre. An optimal extraction (Sharp & Birchall 2010) is performed to fit the flux amplitudes perpendicular to the dispersion axis. Gaussian profiles are fit, holding the centre and width constant (based on the tramline and fibre width maps measured above) and fitting all 819 fibres simultaneously. At the same time a B-spline is fit to model the smooth scattered light. This is typically eighth order, so that a total of 827 parameters are fit at once for each CCD column (4096 pixels). To enforce smoothness on the scattered light model, the fit is done in two passes. On the first pass all 819 fibres plus 8 scattered light parameters are fit (as outlined in Sharp & Birchall 2010). Then the scattered light model is smoothed across columns and the resulting 2D image is subtracted from the data frame. The second pass is then done on the subtracted frame, fitting only the 819 fibre amplitudes, without any scattered light model. An example extraction fit is shown in Fig. 4.

Following extraction, the 1D spectra are divided by an extracted and normalized 1D fibre flat-field spectrum, which removes residual fibre-to-fibre variations in spectral response. These do not correct for total throughput as the illumination is not sufficiently uniform across the flat-field and so no transmission calibration between the fibres is possible at this stage (Sharp, Brough & Cannon 2013). Wavelength calibration is performed using standard arc-lamps (CuAr). Emission lines are identified in extracted 1D spectra and matched to line lists with a third order polynomial for each fibre solution. A secondary wavelength calibration is performed in the red arm by measuring the positions of several sky emission lines and fitting a quadratic to the residuals relative to their known wavelengths. This modification corrects for small shifts introduced due to difference

[2] In 2014 April, the cosmetically poor blue AAOmega CCD was replaced with new device largely free from such defects. The continued need for full bias/dark correction will be reconsidered as experience is gained with the new array. The red CCD will be upgraded in 2014 August.

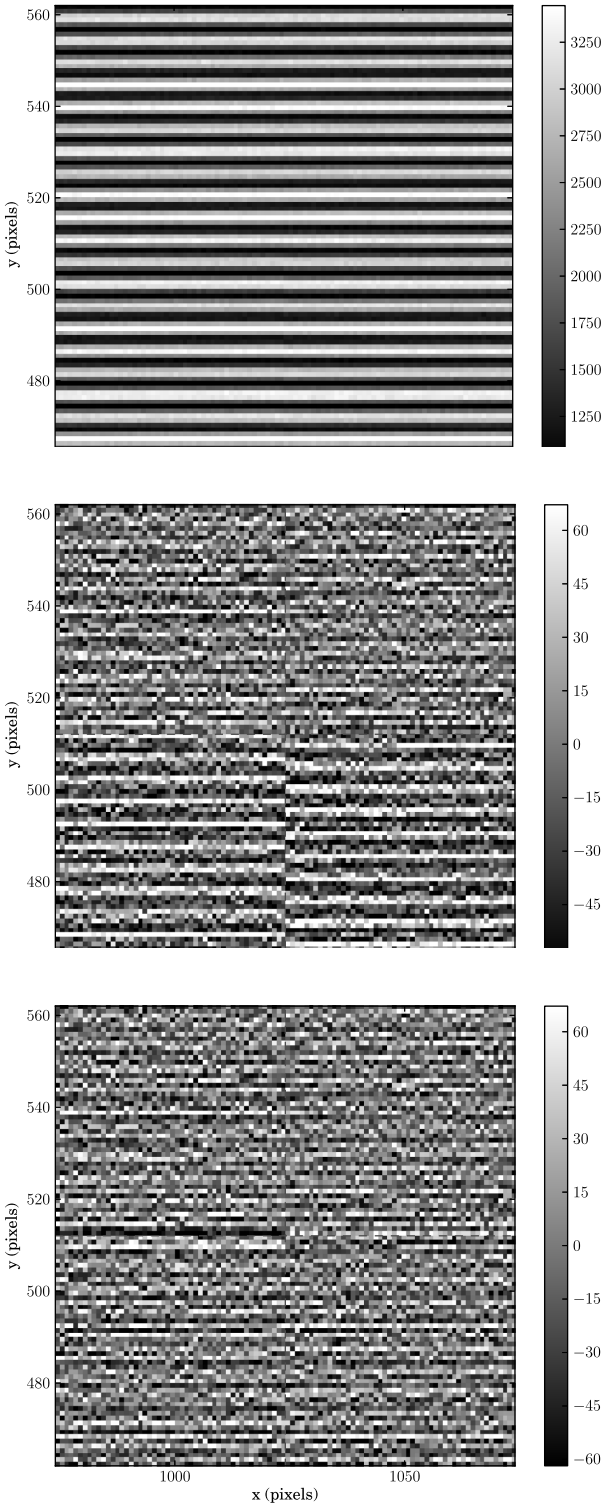

Figure 2. A small region of a SAMI fibre flat-field image (top), along with residuals after extraction (middle and bottom). The residuals are shown without (middle) and with (bottom) correcting for lithography errors. The red dashed lines indicate the CCD lithography boundaries where we would expect discontinuities in the fibre trace.

in the feed angle of sky and calibration illumination. It is found to improve sky subtraction accuracy while not significantly modifying the wavelength solution. This correction cannot be applied in the blue arm, where only the 5577 Å sky line is in the spectral range observed.

The relative throughput between all fibres is established via measurement of several isolated night sky emission lines. The requirements for relative transmission calibration are outlined by Sharp et al. (2013). In occasional cases the sky lines are affected by cosmic rays and other artefacts. To minimize the effect of this issue, we use the median throughput value for each fibre across all observations of a field on a single night.

The 26 individual sky fibres within the SAMI spectrograph slit are configured to blank-sky positions in each observation field. Once throughput calibrated, a master sky spectrum is generated by stacking the individual sky fibres with each frame before subtracting the sky. The Principal Components Analysis technique (Sharp & Parkinson 2010) cannot be used due to the common spectral features in most SAMI Galaxy Survey targets, a consequence of the limited redshift range for the survey and the coherence of spectral structure within each galaxy.

The individual fibre spectrum processing with 2DFDR is now complete and the RSS frames are passed to an external processing suite to flux calibrate and correct for differential atmospheric refraction (DAR) and dispersion before aligning and mosaicking individual objects to produce data cubes.

4 FLUX CALIBRATION AND TELLURIC ABSORPTION CORRECTION

The flux calibration for each set of galaxy observations has two parts. First, a spectrophotometric standard star – typically observed on the same night as the galaxy observations – is used to correct for the large-scale wavelength dependence of the instrumental transmission. A secondary standard star – observed simultaneously with the galaxies – is then used to correct the telluric absorption bands. A full analysis of the flux calibration accuracy achieved for the SAMI Galaxy Survey Early Data Release is given by Allen et al. (2015). Below we briefly describe the principles adopted.

4.1 Extracting total stellar spectra

In principle, each stage in the flux calibration can be performed by multiplying the observed galaxy spectra by the ratio of the true flux to the observed flux of a standard star. For IFU observations with a filling factor less than 100 per cent, we must also account for the light that falls between the fibres. Crucially, the fraction of light lost in this manner is a function of wavelength, as a result of atmospheric dispersion and the dependence of seeing on wavelength. We make the correction by fitting the observed stellar flux across the hexabundle with a model point spread function (PSF), including the atmospheric effects. The model PSF as a function of x, y position (in arcseconds) and wavelength, λ, takes the form of a Moffat (1969) profile:

$$p(x, y, \lambda | x_{\rm ref}, y_{\rm ref}, \lambda_{\rm ref}, \alpha_{\rm ref}, \beta) = \frac{\beta - 1}{\pi \alpha(\lambda)} \left(1 + \left(\frac{(x - x_0(\lambda))^2 + (y - y_0(\lambda))^2}{(\alpha(\lambda))^2}\right)\right)^{-\beta}, \quad (1)$$

where $x_{\rm ref}$, $y_{\rm ref}$ and $\alpha_{\rm ref}$ give the position and size at an arbitrary reference wavelength $\lambda_{\rm ref}$. The dependence on wavelength is given

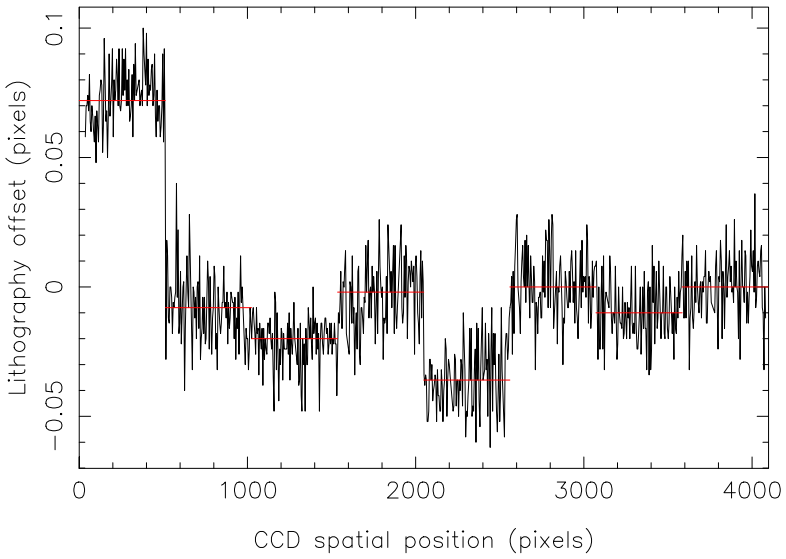

Figure 3. CCD lithography errors measured during fibre tracing. The black line shows the offsets measured for each fibre and the red lines show the median measured offset in each 512 pixel block.

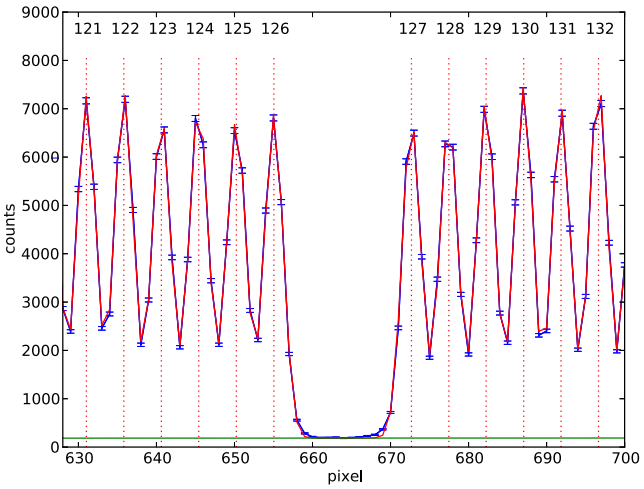

Figure 4. Example fibre profile data (blue line with errors) from a fibre flat-field frame. Overplotted (red) is the best-fitting multi-Gaussian model from our optimal extraction. The smooth scattered light model is shown in green. The vertical dashed lines show the expected centres of each fibre from the tramline mapping process, with fibre number along the slit labelled above. The gap at pixel ≃665 between fibres 126 and 127 is due to the separation between individual slitlets each containing 63 fibres in the SAMI slit.

by

$$x_0(\lambda) = x_{\rm ref} + 206250\,(n(\lambda) - n(\lambda_{\rm ref}))\tan(ZD)\sin(\phi),$$
$$y_0(\lambda) = y_{\rm ref} + 206250\,(n(\lambda) - n(\lambda_{\rm ref}))\tan(ZD)\cos(\phi),$$
$$\alpha(\lambda) = \left(\frac{\lambda}{\lambda_{\rm ref}}\right)^{-0.2}\alpha_{\rm ref}, \qquad (2)$$

where ZD is the zenith distance and ϕ is the parallactic angle. The refractive index of air as a function of wavelength, $n(\lambda)$, is given by equations 1– 3 of Filippenko (1982), which are in turn functions of temperature, pressure and water vapour pressure.

To constrain the free parameters, the observed wavelength range in each CCD is divided into 20 chunks of 100 pixels each, having discarded the 24 pixels at the beginning and end of each CCD. The observed counts in each fibre are summed within each of these wavelength ranges, and the model PSF is fitted to this summed data. The data from both CCDs are fitted simultaneously, in order to have a wide wavelength range to constrain the atmospheric dispersion effects. During this fit, the parameters $x_{\rm ref}$, $y_{\rm ref}$, $\alpha_{\rm ref}$, β and ZD are allowed to vary along with the total flux in each wavelength range, while ϕ and the atmospheric parameters are fixed to their measured values.[3] For SAMI observations under typical observing conditions the uncorrected atmospheric dispersion between 3900 and 7270 Å is of the order of 1 arcsec, ∼60 per cent of a fibre core diameter. After correction the mean offset between these two wavelengths, averaged over a series of observations, is found to be 0.14 arcsec. At less than 10 per cent of the fibre core diameter, this is deemed to be at the limit of measurement.

With the PSF parameters set, the final step in extracting the spectrum is to fit for the overall flux, i.e. the scaling of the PSF, in each wavelength pixel. A uniform background level is also fit in each wavelength pixel, to allow for residual errors in the sky subtraction. Each wavelength pixel is fit independently. The result is a spectrum recording the total number of CCD counts that would have been observed, if the filling factor of the IFU was 100 per cent.

4.2 Primary flux calibration

The spectrum for the primary standard (typically multiple standard stars are observed, at a range of airmasses) is then corrected for atmospheric extinction using the default Siding Spring Observatory (SSO) extinction curve (scaled for airmass) and rebinned to match the wavelength sampling of the reference spectrum, given in units of erg s^{-1} cm^{-2} Å^{-1}. The ratio of the observed and reference spectra is then used to infer the wavelength-dependent scaling needed to calibrate the data in an absolute sense, or, equivalently, the instrumental response function, $R(\lambda)$.

To minimize the effect of noise in the standard star observations, and of small-scale mismatches between the observed and template

[3] The zenith distance dependence is largely degenerate with atmospheric pressure. Fitting accuracy was found to improve marginally if minor variations in ZD were allowed as a free parameter of the model.

spectra, the measured ratio is smoothed by fitting a spline function. Approximately eight spline knots are used within each arm, although extra knots are inserted around 5500 Å where $R(\lambda)$ shows a sharp turnover due to the dichroic cut off, and knots that lie within telluric bands are removed.

After smoothing, multiple observations of each standard star from a given night are combined to produce the final calibration. The agreement between different observations is very good in terms of the shape of $R(\lambda)$, but there is some variation in the overall scaling with a standard deviation of 11.1 per cent found in the available year-1 data set. This variation is driven by a combination of changes in atmospheric transmission and a weak degeneracy between the scaling and the full width at half-maximum (FWHM) in fitting the PSF. To remove the effect of any outliers with discrepant normalizations, the individual measurements of $R(\lambda)$ in each arm are re-scaled such that their value in the centre of the wavelength range is equal to the median of these values across the set of observations. The rescaled measurements of $R(\lambda)$ are then combined using a simple mean. The response function is then applied to each science frame after correction with the airmass-scaled SSO extinction curve. Details of the accuracy of the calibration process, with SAMI data cubes referenced to ancillary data (e.g. SDSS photometry and spectroscopy) are given for the SAMI Galaxy Survey Early Data Release by Allen et al. (2015).

4.3 Telluric absorption correction

The Fraunhofer *B* band, a telluric absorption feature due to atmospheric O_2 at 6867 Å, falls within the red portion of the SAMI Galaxy Survey spectra and requires correction. Because the telluric band strength is variable and strongly correlated with airmass, the correction is best derived from data contemporary with the science observations. For every science observation, 1 of the 13 SAMI bundles is allocated to a secondary standard star. These secondary standards are colour selected to be F subdwarfs, which have relatively flat and featureless spectral shapes, allowing simple modelling with either synthetic or empirical template spectra. The total spectrum for each secondary standard is extracted using the process described in Section 4.1, after the RSS data have been flux calibrated as per Section 4.2. The secondary standards are used to correct for telluric absorption. As a byproduct of this process, the secondaries also provide per-data frame information about the seeing, including an empirical measurement of DAR. This information is used in the later alignment and drizzling stages.

The spectrum of a F subdwarf is sufficiently smooth around the regions of telluric absorption (6850–6960 Å and 7130–7360 Å) that it can be modelled with a simple straight line fit. The fit is performed using all data outside of the telluric regions and redwards of Hα. The telluric absorption is then given by the ratio of the extracted spectrum to the straight line fit; all galaxy spectra in the frame are divided by this ratio, with the correction applied only in the vicinity of the telluric feature.

4.4 Absolute flux calibration

The flux calibration procedure assumes no change in atmospheric conditions (other than for the telluric absorption features) between the observations of the spectrophotometric standard star and the galaxy field. In practice, some variation can occur, which to first order produces a wavelength-independent scaling of the observed flux. To correct for this scaling, the spectrum of the secondary standard star is extracted from the combined data cubes using the same procedure as described in Section 4.2, and converted to a *g*-band magnitude by integrating across the SDSS filter curve. This measurement is compared to the available photometry to find the appropriate scaling, which was then applied to all objects in the field.

5 RESAMPLING TO A CARTESIAN GRID

The SAMI hexabundle format imposes two restrictions on image reconstruction which are shown graphically in Fig. 1. While the positions of the individual fibre cores that make up each IFU are well defined (± 1 μm relative errors, $\sim$1 per cent of a fibre diameter), they are on an irregular grid. Secondly, each hexabundle IFU has a 73 per cent fill factor for a single observation. The full image profile is recovered through a conventional series of dithered observation with the telescope offset to fill in the gaps in target coverage. Dithering generates a series of misaligned data frames which possess a well registered but non-common geometry. In principle, one can retain the data in this native format, but the exigencies of scientific analysis typically dictate that data should be rebinned, often to a common Cartesian output grid. This facilitates straightforward stacking of dithered frames to allow outlier rejection and accrual of signal-to-noise ratio (S/N) as well as easy visualization of resulting data products, and significantly simplifies many subsequent analysis steps.

A simple option for resampling is to perform a nearest-neighbour interpolation of the hexabundle input data on to a regular Cartesian grid. An immediate effect of such an operation is to effectively convolve the intrinsic resolution of the spatial data with a kernel corresponding to the chosen output spaxel size, blurring the image. Secondly, such a resampling introduces a complex covariance between output spaxels which overlap more than one input fibre core. Both issues can be minimized by selecting a fine pitch for the output spaxel grid, but this produces a bloated data format, containing significant redundant information, and ultimately reduces S/N due to the excessive oversampling. The problem is well documented, and a solution clearly defined by Fruchter & Hook (2002) through their introduction of the *Drizzle* algorithm, initially conceived to resample high-resolution imagining from *Hubble Space Telescope* (*HST*)/WPFC2, which subcritically samples the PSF from *HST*.

For a detailed description of the drizzle algorithm, the reader is directed to Fruchter & Hook (2002). Briefly, considering each input fibre core in turn, and for any given input geometry, one can calculate the overlap area of the input fibre core with each element of a predefined regular grid of output spaxels. The fractional area of an input fibre core covering each output spaxel dictates how the flux should be redistributed to each output spaxel (Fig. 5). This fractional area provides a *weight* for each output spaxel which represents the relative exposure of each output spaxel.

Output spaxels that do not sit within any input fibre cores will be assigned a weight of zero. Output spaxels that fall on the border between multiple input cores will have a weighted contribution from each core, but the total weight for such spaxels will not exceed unity (since by definition, an output element cannot be completely inside one input element if it is also part of other input elements). To place dithered data on to the regularized grid, one merely recalculates the overlaps after perturbing the baseline position (by the known telescope offset) of the input fibre cores relative to the initial reference position. Fig. 6 presents an example of the relative weight mappings for three observation that make up part of a dither set.

For isolated output spaxels, this process conserves total flux but does introduce covariance, an issue we return to in Section 5.6. In

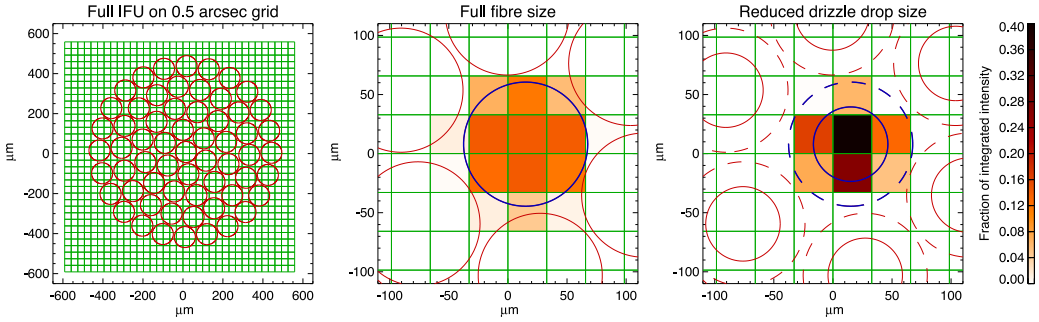

Figure 5. An example of the SAMI implementation of the drizzle algorithm (Fruchter & Hook 2002) shows the redistribution of flux for a SAMI IFU bundle on to a regular Cartesian output grid. The 0.5 arcsec output spaxel grid is chosen to encompass the entire IFU dither pattern for a series of observations (left), the regular output grid is shown below the fibre bundle footprint for a single dither position. For each input fibre core, each output spaxel receives a portion of the input flux (centre). As we show in Section 7, loss of spatial resolution can be minimized by distributing the flux from each input fibre core as if it had a reduced the diameter (right) with the flux distributed over a smaller number of output spaxels and at higher intensity in each.

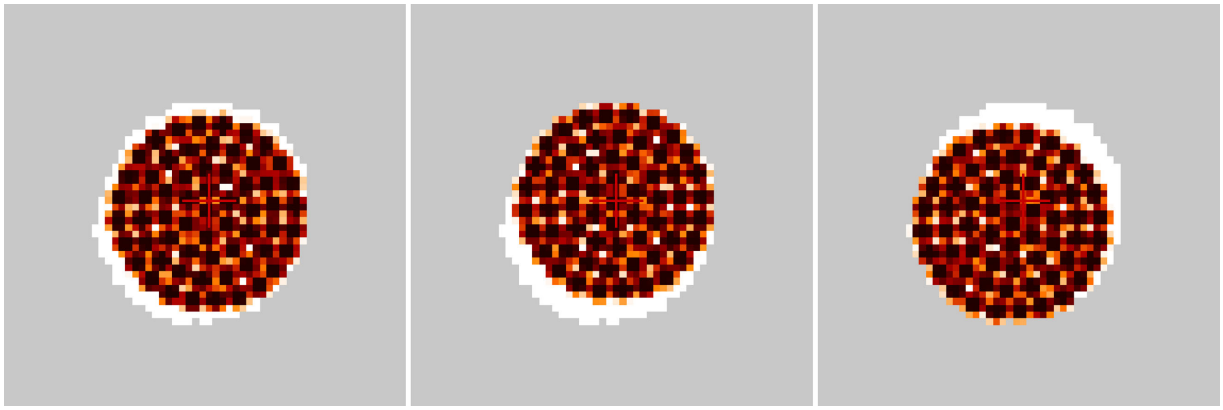

Figure 6. Weight maps are shown for three data cubes from an aligned dithered data set. Grey spaxels have no coverage in the current output geometry, while the white-to-black scaling indicates weights between 0 and 1. The fibre core centres are visible as spaxels with saturated (black) weights. Note, the large white borders to each cube show zones of zero overlap between an individual image and the overall image stack. Weight maps can be interpreted as the relative exposure time of each spaxel of the image. The output spaxel scale is set to 0.5 arcsec which means that while most output spaxels have some fractional contribution from one or more SAMI fibres, some spaxels inside individual cubes have no contributing data.

this regime, the associated error information for each spaxel can be redistributed such that the global S/N is preserved, i.e. the input variance is simply weighted by the square of the weight map.

5.1 Weight cubes

The weight maps derived from the drizzle resampling trace the effective exposure time for each output spaxel. By design, not all output spaxels have the same effective exposure time. Those that are not completely covered by one or more IFU fibre cores will have a reduced intensity proportional to the fractional area of the output spaxel covered by input fibre cores. Depending on the output spaxel size chosen, some spaxels may have an effective exposure time of near zero even though they reside within the physical boundary of the IFU bundle. For this reason, a reconstructed image of a source within the drizzle resampled data cube will exhibit an intensity structure dominated by the relative weights of its component spaxels, largely obscuring the underlying source structure.

To overcome this effect, the resampled data cubes (and their associated variance arrays) are stored with the weight map normalization pre-applied (i.e. mosaic cubes are divided by their weight map in the default data product). This normalizes the effective exposure time across the data cube and recovers the underlying source structure. This data format means most users may safely ignore the weight maps for most applications. Provided the weight map is propagated alongside each data frame, the process of conversion between the two formats is reversible and deterministic. Examples of three weight maps for a three-point dithered observation are shown in Fig. 6 and the imprint of variations in relative exposure across a resampled observation is shown in Fig. 7.

As dithered observations are combined, the relative weighting of all spaxels will approach uniformity across the mosaic and the relative exposure structure will be removed from the data cube. The practical implication of this is as follows.

(a) *Local properties* of a source, such as surface brightness of emission features or light profile fitting, should use the default data format with the weighting already applied to the data cubes.

(b) *Global properties* of a source, such as integrated Hα flux or continuum intensity, should pay strict attention to the weight map values to avoid erroneously including additional signal, particularly from the edges of the mosaic, which are highly scaled up from low relative exposure times with respect to the central regions of the cube.

When combining data, cubes are first multiplied by their weight maps, the data and individual weight maps summed, and the summed data finally divided by the new weight map to preserve the correct flux calibration.

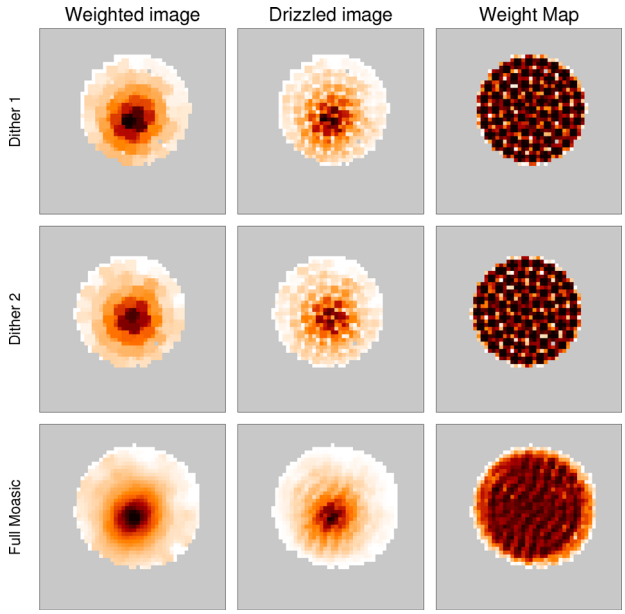

Figure 7. The effect of the weight map is illustrated with three mosaics of a SAMI galaxy. The two upper rows each show a single dither position observation for the galaxy. The bottom row shows the reconstructed image after seven dither positions are combined. The first column shows the reconstructed image after division by the weight map, showing a constant effective exposure time for all spaxels. The second column shows the raw output from the drizzle process where the intensity distribution is heavily modulated by the different effective exposure times of each output spaxel due to the partial coverage by input fibre core. The third column shows the weight map, the effective exposure time for each spaxel.

5.2 Dither frame alignment

The relative offsets between each dithered data set must be known before they can be drizzled to a common output spaxel grid and combined. In principle, these offsets can be obtained from telescope pointing and offset information available in each RSS frame header. In practise, this information does not provide the required level of accuracy, particularly when data are taken with multiple source acquisitions (for example over multiple nights), a process which can affect the base pointing position.

A number of methods were explored for alignment, largely treating each IFU individually. These included a simple centroid fitting, cross-correlation of reconstructed Integral Field Spectroscopy (IFS) images and alignment using SDSS imaging data as a reference frame. It was found that all work remarkably well for most source types, but fail significantly for some classes of objects, such as galaxies with disturbed morphologies, interacting pairs or low surface brightness systems.

In order to overcome this limitation, and to recover all the observed targets in a uniform manner, an alignment technique has been developed to simultaneously estimate the dither patterns for all IFUs in a given observation. The principle treats each exposure as an image of the sky, and estimates the best-fitting coordinate transformation to align each RSS frame to a reference RSS frame (typically the first observation in a sequence). In this way, even if one (or more) IFUs cannot be used for the estimate of the coordinate transformation (for example due to a disturbed morphology system which provides unstable alignment results), the best-fitting solution allows us to recover the dither pattern for the entire frame. This goal is achieved in three steps.

First, a 2D Gaussian is fitted to an intensity map for each IFU (obtained by collapsing the cube along the wavelength axis) in order to recover the centroid (i.e. peak signal) positions. Assuming that the absolute position of the peak emission on the sky does not vary between exposures, the centroid coordinates can be used to estimate the best-fitting coordinate transformation necessary to align all IFUs on a reference frame.

Secondly, the best-fitting coordinate transformation is computed using a PYTHON implementation of the IRAF[4] GEOMAP task. We allow for a combination of shifts in the x and y directions, a rotation and a common plate scale change in the x and y directions. Specifically, the coordinate transformation has the following functional form:

$$\begin{aligned} x_{\rm ref} &= x_{\rm shift} + (\Delta \times x_{\rm in} \cos\theta) + (\Delta \times y_{\rm in} \sin\theta)\,, \\ y_{\rm ref} &= y_{\rm shift} - (\Delta \times x_{\rm in} \sin\theta) + (\Delta \times y_{\rm in} \cos\theta)\,, \end{aligned} \tag{3}$$

where $x_{\rm ref}$ and $y_{\rm ref}$ are the centroid positions in the reference frame, $x_{\rm in}$ and $y_{\rm in}$ are the centroid positions of the plate to be aligned, $x_{\rm shift}$ and $y_{\rm shift}$ are the rigid shifts in the x and y directions, Δ is the magnification factor and θ is the rotation angle. A 2σ clipping technique is applied to remove those IFUs (3 on average) for which the 2D Gaussian fit is unstable. The mean values of shift, magnification and rotation angle found for our data are $\sim$35 μm, $\sim 10^{-4}$ and $\sim 0.014^{\circ}$, respectively.

Thirdly, we use an implementation of the IRAF GEOXYTRAN task to apply the coordinate transformation to the central fibre of each IFU and determine its position on the reference plate. For each IFU the relative offset between the two exposures is then given by the difference in the positions of the central fibres. While the rotation term introduces a significant translation in base [x,y] position between observations, the magnitude is sufficiently small that no accounting is made for rotation of fibre positions within a bundle (the correction for internal rotation would be typically of the order 0.2 μm, less than 1 per cent of a fibre core diameter and smaller than the relative positional uncertainty of fibre cores within each bundle).

In data analysed to date, the typical rms for the final dither solution is $\sim$12 μm $\pm$8 μm (i.e. 1/9th of the fibre size). These values are found to be consistent with those obtained with the single IFU methods for stable cases indicating the technique is providing a high accuracy and stable solution for all IFUs across a given SAMI observation.

5.3 Simple summation combination

Once relative alignment has been determined, dithered observation data sets can be combined. We start by assuming n data cubes are available from a dithered set. Each data cube C_n is an array of [x,y] spaxels (which for compactness we shall denote by r, with the assumption that r is defined in a reference frame with all cubes aligned). We also assume, as is the default for SAMI data, that the data cubes are stored with division by the weight map pre-applied (Section 5.1). Each data cube also has λ wavelength elements that, due to atmospheric dispersion, are offset with respect to one and other. This misalignment is corrected by recomputing the drizzle mapping solution as a function of wavelength at regular intervals.

[4] IRAF is distributed by the National Optical Astronomy Observatory, which is operated by the Association of Universities for Research in Astronomy (under cooperative agreement with the National Science Foundation.

The interval is chosen such that the accumulated dispersion misalignment is never more than 10th of a spaxel. Data cubes need not be a common size in [*x*,*y*], although the spaxel scale of the cubes to be combined must be constant[5] and the region of union between cubes well defined. Each cube has an associated variance map V_n and map of the relative spaxel weights W_n.

We define an output data cube, $C_{\rm out}$, and its associated variance array, $V_{\rm out}$ and weight map, $W_{\rm out}$, such that

$$W_{\rm out}(r,\lambda) = \sum_n W_n(r,\lambda), \tag{4}$$

$$C_{\rm out}(r,\lambda) = \frac{1}{W_{\rm out}(r,\lambda)} \sum_n C_n(r,\lambda) W_n(r,\lambda), \tag{5}$$

$$V_{\rm out}(r,\lambda) = \frac{1}{W^2_{\rm out}(r,\lambda)} \sum_n V_n(r,\lambda) W^2_n(r,\lambda). \tag{6}$$

The default values of C_n must be multiplied by their weight maps in order to restore the true relative exposure time structure to the input data before summation. After summation, the output cube is divided by the new weight map to remove the relative exposure time structure from the final data product. Not applying the weight map would leave the cube scaled for differences in the effective exposure times for each spaxel and would not correctly propagate observed flux. For many spaxels the weight will eventually exceed unity as multiple exposures are summed together. This merely indicates that the spaxel is fully sampled at a higher S/N than would be achieved for unit exposure time, due to multiple overlapping observations. While these relative exposure values must be carefully propagated in the image header, they present few problems. Arbitrary renormalization of the weight map is permissible, provided the effective exposure time of the associated data cube is also renormalized by the same factor. This preserves the correct surface brightness for the data cube.

5.4 Outlier rejection

Cosmic rays and other defects will remain in the reduced data cubes at some level. Additionally, bad pixels from the CCD will contain no data and hence will be missing/flagged in the data cubes. Clipping flagged values is as simple as removing them from summation and reducing the corresponding value of N for each (r, λ) in equation (4). The key is accurate flagging.

To remove outliers from the summation of pixel values a clipped mean is preferable because it simplifies variance propagation. To generate the clipping flags, we work on each spaxel and spectral pixel of the output cube (r, λ) independently, noting that there will be N intensity values at each output pixel, one from each input data cube. Note also that some input values will have weight $W_n(r,\lambda) = 0$ due to the nature of the dithered observation. We generate the working vector, $m(n)$, and its associated variance array such that

$$m(n) = [C_0(r,\lambda|W_0 \neq 0), \ldots, C_n(r,\lambda|W_n \neq 0)]. \tag{7}$$

At this point a simple sigma-clipping rejection, with a modest threshold, will fail due to the finite fibre footprints which sample different parts of each source due to the dithered observation strategy. Put simply, in comparing the dithered input spectra at each spaxel, one is not directly comparing like-with-like.

[5] In principle the inputs to the drizzle resampling step need not all be at a common scale, provided the correct geometry is supplied to the drizzle code.

For all real sources, which lack spatial discontinuities on the scale of SAMI data due to the seeing profile, the first order difference will simply be intensity. There will be no appreciable change in spectral shape across the output spaxels. Therefore, if each input spectrum is first normalized to unity prior to construction of the vector $m(n)$ the sigma-clipping outlier rejection flags only significantly deviant pixels (a 5σ-clipping threshold is used). With the errant pixels now cleanly identified, the remaining good pixels are used, without normalization, in the summation equations of equation (4). In this manner, highly discrepant data points are removed from the summation, while still retaining the correct intensity information from the dithered data set.

5.5 Confirming variance propagation accuracy

The default data product presented above generates a variance array to track S/N for each input fibre spectrum as it is added to the dithered output mosaic. This simple data product fails to account for the significant correlation introduced between adjacent spaxels. By definition the covariance matrix is given by

$$\Sigma(i,j) = \mathbb{E}\left[(x_i - \mathbb{E}[x_i]) \cdot (x_j - \mathbb{E}[x_j])\right], \tag{8}$$

where $\mathbb{E}(x_i)$ is the expectation value of the data in spaxel i. Consider two adjacent output spaxels of the data cube, i and j which sit beneath a common input fibre core with $I_0 \pm \sigma_0$ that contributes a fraction of flux to each output spaxel α_i and α_j. Then $\mathbb{E}(x_i - \mathbb{E}(x_i)) = \alpha_i \sigma_i$ and hence the variance term for spaxel i is $\Sigma(i,i) = \sigma_i^2 = \alpha_i^2 \sigma_0^2$ and the covariance between output spaxels is $\Sigma(i,j) = \alpha_i \alpha_j \sigma_0^2$.

In order to test this error propagation model, a simulated data set is generated. A dithered sequence of noise-free image frames is generated and drizzled on to the standard output grid. The input data are then duplicated and a Gaussian random noise field of known distribution is added. This second data set is also drizzled on to the standard output grid. If the associated noise properties of the image are correctly propagated by the error model, then a histogram of the difference between pixel intensity values for the noise-free and noisy data, scaled by the error array, will be a Gaussian distribution centred on zero with unit width. This is confirmed to be the case, indicating accurate propagation of variance information from RSS frames to SAMI data cubes. This analysis confirms the correct propagation of the statistical errors recorded during the data reduction process. Underlying systematic variations in the data cubes, due principally to variations in observing conditions during survey operations, are assessed by Allen et al. (2015).

In essence, the drizzle process creates a number of identical copies of each input spectrum, each copy scaled by a value α_i. Since duplication and scaling of a spectrum must retain the intrinsic noise properties, the associated error array simply scaled by α_i to maintain the S/N in each copy of the spectrum. The noise properties of each output spaxel are correctly traced, but this simple model has made no accounting for the covariance between output spaxels, it merely correctly retains the noise properties of individual input spectra.

5.6 Neglecting covariance

Neglecting covariance between spaxels makes the assumption $\sigma_{ij}^2 = 0$ for $i \neq j$. With this assumption, on summation of the output error values for a dithered data set we no longer recover the input error, σ_0. As the α_i (with $0 \leq \alpha_i \leq 1$) sum to unity, $\sum(\alpha_i) = 1$, the quadrature sum is generally less than unity, $\sum(\alpha_i^2) \leq 1$. The S/N is correctly modelled within each individual output spectrum, but the

noise level is underestimated across the full mosaic due to the lost covariance information. While a crude fix for this approximation would be to rescale all the σ_i^2 by the factor $1/\sum(\alpha_i^2)$, a method for tracking the covariance is considered below.

5.7 Tracking covariance

The drizzle resampling methodology introduces unavoidable covariance between the output spaxels. For a detailed description of the problem, see section 7.1 of Fruchter & Hook (2002). In principle the covariance information can be tracked for each output pixel. Indeed, the diagonal elements of the full covariance matrix (i.e. the variance) are already tracked in full. For an input fibre core diameter of 1.6 arcsec, and an output spaxel grid pitch of 0.5 arcsec, there are ∼16 output spaxels with non-zero covariance for each input fibre core (although only 4–9 of these have significant covariance depending on the specific input/output geometry). On mosaicking dithered data of this format each output spaxel has a non-zero covariance only within a 5 × 5 spaxel grid of adjacent spaxels when adopting the default SAMI Galaxy Survey observing strategy (a seven-point dithered mosaic with a dither pitch of 0.72 arcsec, to be considered in Section 6, on to 0.5 arcsec spaxels with a 50 per cent drizzle drop size reduction to be presented in Section 7). Explicit evaluation of the covariance between spaxels confirms there is no covariance outside of the 5 × 5 spaxel grid. However, even assuming these limited overlaps, providing the full covariance information in the output mosaic files would require a significant and largely unwarranted increase in the data volume and processing time for each observation.

When calculating and retaining the full covariance information, a data cube of ∼1024 Mb is generated, requiring a runtime of ∼100 min on a standard dual core desktop machine – a data volume and processing time that are prohibitively high for a large galaxy survey such as the SAMI Galaxy Survey. Both data volume and processing time can be significantly reduced by noting that the structure of the 5 × 5 covariance maps varies slowly with wavelength. By sampling at regular intervals in wavelength space and then interpolating between the measured values along the wavelength axis the full covariance matrix can be recovered with only minimal loss of information. The covariance cubes generated in this way are stored normalized to the variance (i.e. the central pixel of each 5 × 5 covariance map has value 1.0) and require scaling by the variance to recover the correct magnitude. Furthermore, the covariance maps are generated from the overlap fractions of the input fibre footprints, and hence apply the flux and variance values before normalization for relative exposure time with the weight maps.

As an example, consider the data cube, $C[x, y, \lambda]$, and its associated variance and weight arrays, $V[x, y, \lambda]$ and $W[x, y, \lambda]$. Two adjacent spaxels, A and B, will be covariant. The data cube contains the normalized spaxel values $C[x_A, y_A, \lambda]$ and $C[x_B, y_B, \lambda]$, each of which has an input flux and variance, prior to normalization by the weight map, given by

$$C'[x_n, y_n, \lambda] = C[x_n, y_n, \lambda] \times W[x_n, y_n, \lambda] \quad (9)$$

$$V'[x_n, y_n, \lambda] = V[x_n, y_n, \lambda] \times W[x_n, y_n, \lambda]^2. \quad (10)$$

The covariance of spaxel A with spaxel B is contained within the covariance matrix, *Covar*, such that the covariance between $C'[x_A, y_A, \lambda]$ and $C'[x_B, y_B, \lambda]$ is given by

$$Covar[x_A, y_A, x_B, y_B, \lambda] \times V[x_A, y_A, \lambda] \times W[x_A, y_A, \lambda]^2. \quad (11)$$

In addition to sampling at regular wavelength intervals, it is critical to finely sample the full covariance matrix (in the wavelength direction) either side of wavelength slices at which the applied atmospheric dispersion correction changes. The correction shifts the on-sky positions of the input fibre cores relative to their position on the CCD, and therefore alters the covariance between output spaxels. As a compromise between adequately sampling the full covariance matrix, and minimizing the processing time required to produce a data cube and the volume required to store that cube we opted to sample the covariance matrix at 100 pixel intervals along the wavelength axis, and additionally over a region of ±2 pixels at each wavelength slice the applied atmospheric dispersion correction is updated. This results in data processing times of ∼10 min and data cube volumes of ∼170 Mb. The full covariance matrix is then recovered by interpolating between wavelength slices with covariance information, with the full covariance matrix produced in this way recovering >80 per cent of the true covariance (Fig. 8). The SAMI data cubes contain an additional `COVAR` extension which contains the covariance information sampled in this fashion. The `.fits` header of this extension records the wavelength slices (in pixels) at which the covariance was sampled, with the header item, `COVARLOCn` giving the wavelength slice of the *n*th element along the wavelength axis of the sparsely sampled covariance matrix. An example implementation of the simple interpolation necessary to recover the full data is found in Appendix A.

The effectiveness of this covariance compression model is assessed by comparing the full covariance matrix to the modelled data on a spaxel by spaxel basis. Each spaxel is assessed across the 5 × 5 covariance grid, normalized to unity centred on the spaxel (i.e. for unit variance). The difference between the true covariance matrix and the modelled data is then summed, both while retaining the sign of differences and also as an absolute sum of differences. A cumulative histogram of these sums is then generated (Fig. 8). It is found that for 85 per cent of spaxels the missing covariance signal amounts to <10 per cent the variance associated with each spaxel for the simple summation. When considering the summed absolute differences the missing covariance signal remains <20 per cent of the variance for 80 per cent of spaxels. This level of accuracy is considered to be sufficient for most SAMI Galaxy Survey purposes. For specific individual cases in which higher accuracy is required, individual galaxy data cubes will be reformed with either finer sampling or full covariance array generation.

6 OPTIMAL DITHER STRATEGY

For a given fibre geometry within each SAMI IFU, and for a given number of dithered observations, it is necessary to determine the optimal dither pattern which provides the *most uniform coverage* of the output image map. In essence, we wish to provide all parts of the combined output image with a uniform effective exposure time, i.e. we wish to have a uniform weight map. The metric developed to optimize the alignment is to minimize the ratio of the standard deviation of flux weights in the output cube relative to the median flux weight.

The minimization is restricted to a circular region which completely encompasses the *central* position of the dither pattern (typically the first observations).

Uniform coverage is trivially achieved by increasing the number of dither positions available. This is of course limited by finite observation time, CCD readout overheads, and the sensitivity requirement of achieving sky-limited dither exposures, all of which restricts the number of independent frames. Analysis reveals that

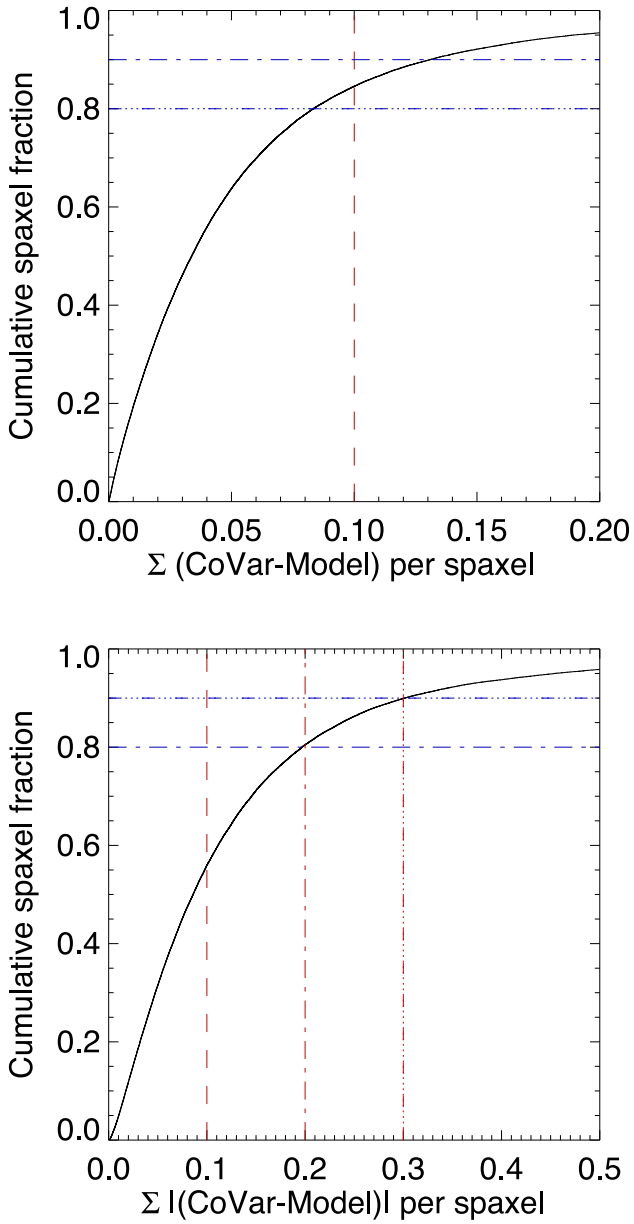

Figure 8. The difference between the true covariance matrix and the modelled matrix is presented as the cumulative histogram of spaxels. The upper figure presents the cumulative histogram summing all underestimates and overestimates in the difference between the covariance model and the full matrix. The lower histogram considers the sum of absolute differences.

while the optimal dither strategy lies within an extended plateau of parameter space (i.e. there are many essentially optimal strategies) it is surprisingly simple to select pathologically bad strategies that provide highly suboptimal coverage and structure with a wide range of effective exposures times (or even image holes) across the final mosaic.

For the fibre arrangement of the SAMI hexabundles, a seven-point hexagonal close-packed strategy, with a radial dither offset of 45 per cent of the fibre core diameter (i.e. 0.72 arcsec for the 1.6 arcsec SAMI fibres) was selected as the default SAMI observation mode (Fig. 9). The impact of the rotation angle for this pattern was explored, noting that each SAMI hexabundle has an accurately known but distinct *lattice* structure. The effect was found to be negligible due to the high fill factor of the SAMI IFS, the optimization being dominated by local coverage around each fibre core. For a system with lower fill factor, the orientation of the offset pattern relative to the fibre-core lattice would likely be significant.

The most unexpected outcome was the realization that the dither pitch should be less than half the fibre-core diameter in order to prevent the appearance of significant coverage holes at the centre of each core position. A further complication is the imprint of the dither strategy on the resolution of stacked data cubes. This issue is addressed in the next section.

7 RECOVERING RESOLUTION VIA REDUCED DRIZZLE DROP SIZE

The main motivation for the original drizzle algorithm (Fruchter & Hook 2002) was to provide a means of recovering resolution in undersampled *HST*/WFPC2 images and avoid introducing a convolution with the selected output pixel size which would further degrade the undersampled PSF. Using a number of subpixel dithered observations of the input image, Fruchter & Hook (2002) demonstrate that not only can one avoid significant degradation of the input image, but also that a significant fraction of the intrinsic resolution of the imaging system can be recovered despite the subcritical sampling of each individual component frame. This is achieved by reducing the effective size of each input spaxel while scaling the flux to account for the reduced area (Fig. 5). This reduction in the drizzle *drop size* results in a smaller footprint for each input spaxels on the output grid. Each additional frame is drizzled on to the output grid but shifted by a fraction of the input spaxel true size and, with careful selection of the drop-size reduction, the process has been shown to result in high-fidelity images which faithfully represent the input image profile while minimizing image degradation due to pixelization by the observing system.

Drizzle is now routinely applied to imaging data from numerous sources and the soundness of the resampling has been confirmed a number of times (e.g. Koekemoer et al. 2011, and references therein). However, in such cases the transformation has always been for a continuously sampled image. Below we explore the effects of drizzling with a reduced drop size on the 73 per cent fill factor and undersampled SAMI data (1.6 arcsec fibres sampling with typical seeing of ∼2.0 arcsec).

7.1 Model data for drizzle resampling

In order to test the impact of the SAMI sampling on recovery of an image PSF, a series of model observations were created. Noiseless stellar profiles, modelled in the first instance as a 2D symmetric Gaussian profile, were generated assuming a seven-point hexagonal close-packed dither strategy. The model data were generated as RSS frames and then mosaicked into data cubes using the SAMI drizzle code. The RSS spectra were generated via numerical integration of the 2D PSF across the face of a model SAMI hexabundle using measured core position properties.

7.2 Recovered PSF width

In all cases a seven-point hexagonal dither was considered, a choice driven largely by the need to reach a final integration time total in individually sky-limited observations. A number of different drizzle drop sizes are also explored for resampling the simulated RSS data to construct the data cube.

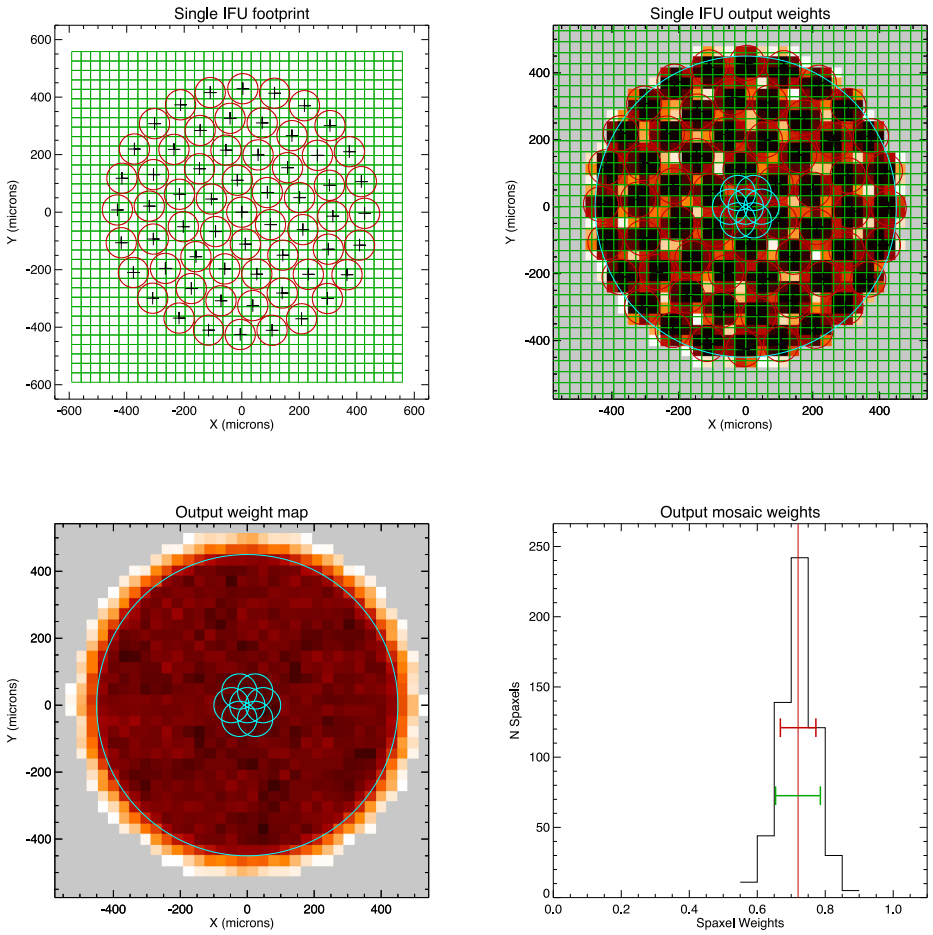

Figure 9. The SAMI data format and dither strategy. Top left: a single hexabundle is shown overlaid on a 0.5 arcsec regular Cartesian grid. Each small circle represents an individual fibre within the IFU bundle. Top right: a drizzle remapping showing the relative weight of each output spaxel for a single observation. The hexagonal dither pattern for a single fibre core is also shown. Bottom left: the relative weight map for a seven-point hexagonal dither. The footprint of a single fibre in the dither pattern is shown overlaid. Bottom right: the histogram of relative weights for the simulated observation. The standard deviation (red) and interquartile range (green) are indicated.

It is apparent from Fig. 10 that using the full drop size generates a broadened profile with a reduced peak intensity. The data are degraded by an amount consistent with the smoothing by the fibre diameter. On reducing the drop size from 1.6 to 0.8 arcsec resolution is recovered although not to the full input image resolution. The image degradation, an inevitable consequence of the subcritical sampling of the input image by the fibre footprint, is dependent on the input image FWHM. For the median seeing expected in survey quality SAMI data, this image blur results in an increase of the image FWHM of ∼0.2 arcsec in simulated data. Reducing the drizzle drop size below 50 per cent of the fibre size does result in a more compact image, but the incomplete sampling due to the limited fill factor of the resulting seven-point dither pattern introduces artificial structure to the image.

7.3 Survey data image quality

A limited quantity of observational data is available with which to confirm the simulated observations. Each SAMI Galaxy Survey plate contains a single calibration star. These standard stars are analysed in the same manner as the simulated data, with the observational seeing assessed in each individual RSS frame via forward modelling of the input stars and the final output image seeing measured directly from each stars dither combined data cube. Data cubes are generated with an output spaxel size of 0.5 arcsec, with a drizzle drop size of 0.8 arcsec, a 50 per cent reduction. At the time of writing, 242 individual RSS frames were available from 36 individual SAMI Galaxy Survey fields. This includes observations over a wide range of observational conditions including some periods of poor seeing. Data taken in the poorest conditions will not comprise part of the final SAMI Galaxy Survey, but are included in this analysis for completeness.

Fig. 10 compares the final recovered seeing for each star to the median of the input seeing for each data set. A chromatic term is visible between the red and blue data cubes, with the blue data typically 0.2 arcsec poorer. This is consistent with the typical wavelength dependence of seeing over the two wavelength ranges. A minor AAT tracking error (identified through analysis of the year-1 SAMI data) is found to introduce a 10 per cent ellipticity in the year-1 SAMI data set. Remedial work is underway at the AAT to rectify the issues. FWHM measurements are taken as the geometric mean of a 2D elliptical Gaussian fit to the data. Fitting residuals are improved slightly, leading to a modest reduction in FWHM, if

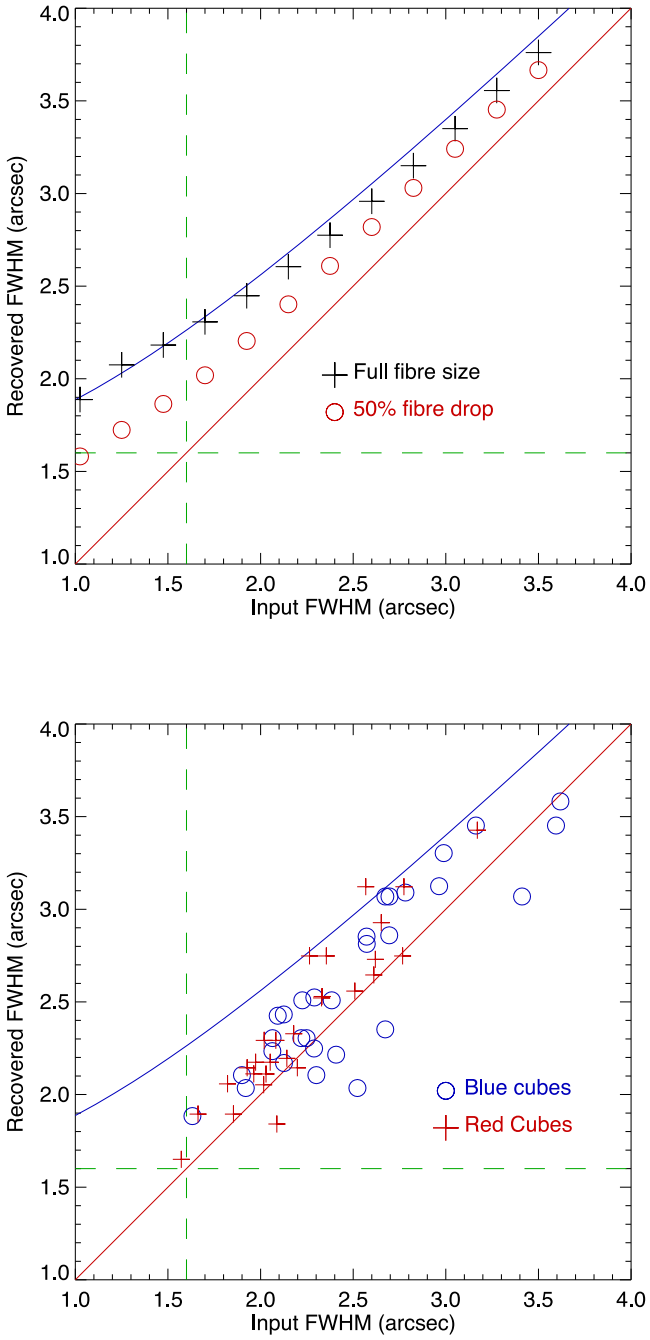

Figure 10. Upper plot: the recovered resolution, with and without a reduction of the drizzle drop size is shown for simulated data. Simulated observations are generated with a range of Gaussian PSF FWHM. The full fibre diameter and a drizzle drop size of 0.8 arcsec diameter (a 50 per cent reduction) are used to resample data on to an output spaxel size of 0.5 arcsec. A diagonal line marks the 1:1 correlation between input and output FWHM, while the curve corresponding to the quadrature sum of the input FWHM and the 1.6 arcsec fibre diameter. The dashed vertical and horizontal lines mark this fibre diameter. Lower plot: for the available standard stars taken from year 1 of the SAMI Galaxy Survey science observations, the median input seeing for each observing block (four to eight dithered frames) is compared with the seeing recovered from fitting to the final data cube of each star.

a 2D Moffat profile (Moffat 1969) is fitted but for consistency with the simulated test data the Gaussian form is presented. The measured cube FWHM values are found to lie above the 1:1 correlation, as expected. The departure from this correlation is not as marked as in the clean simulated data, largely due to the averaging effect for data sets with a marked seeing variation. In early SAMI data products, no explicit resolution matching is performed beyond the simple rejection of markedly discrepant data sets.

The mean and standard deviation of input seeing values are found to be 2.53 arcsec $\pm$ 0.60 arcsec in the blue and 2.25 arcsec $\pm$ 0.43 arcsec in the red. The significant scatter is due to the inclusion of the full year-1 data set in this analysis. For the resulting

cubed data, the corresponding values are 2.61 arcsec ± 0.62 arcsec and 2.40 arcsec ± 0.64 arcsec. This analysis indicates a degradation of the input of not more than 0.2 arcsec due to the alignment and cubing process when performed with the 50 per cent drizzle drop-size reduction. In the context of the input fibre diameter of 1.6 arcsec and typical AAT seeing, this is deemed satisfactory. As the SAMI Galaxy Survey progresses the median seeing of survey data will be improved via judicious flagging and rejection of poor seeing data.

7.4 Flux scaling

Reduction of the drizzle drop size requires a rescaling of the output cube values in order to preserve the flux calibration. Consider a linear reduction in the drop size by a scalefactor ζ (i.e. $\zeta = 0.5$ for a drop size of 0.8 arcsec scaled down from the SAMI fibre core size of 1.6 arcsec). We perform the drizzle resampling of the input flux RSS frame on to an output spaxel grid with the input fibre geometry set to incorporate the reduction factor. This produces the usual set of output data product, $C(r, \lambda)$, $V(r, \lambda)$, & $W(r, \lambda)$. Consider also the intermediate data cube prior to division by the weight map, $C'(r, \lambda) = C(r, \lambda) \times W(r, \lambda)$. If the total flux in the output cubes is calculated for a data set with and without the scalefactor ζ clearly the values will differ by a factor of ζ^2. This is a consequence of the smaller fibre foot prints covering fewer output spaxels and so less flux is distributed. In order to preserve the flux we find

$$\begin{aligned} C' &= C/\zeta^2, \\ V' &= V/\zeta^4, \\ W' &= W/\zeta^2. \end{aligned} \tag{12}$$

However, because the default SAMI data products are provided with the data and variance cubes divided by the weight map, in order to remove relative exposure time variations across the mosaic, the ζ^2 scaling is effectively removed from the output data cube (the scaling being tracked in the weight array, W). This means that for most applications, users need not concern themselves with the weight maps.

8 CONCLUSION

To generate the data products necessary to realize the key science goals of the SAMI Galaxy Survey we have devised a scheme to generate Cartesian gridded data cubes from the irregularly sampled SAMI observational data products. Adapting established image processing techniques, we have demonstrated SAMI data can be accurately regularized, accounting for complications such as atmospheric refraction and dispersion, flux calibration and incomplete sampling of the focal plane. Image degradation due to the sparse sampling and reconstruction is limited to that inherent in the use of an input fibre core comparable in size to the natural image seeing scale. Photometric error information derived during data processing is accurately propagated as well as a compact representation of the complex covariance inevitable in the data. Using these techniques, the SAMI Galaxy Survey is exploring high-multiplex integral field spectroscopy.

ACKNOWLEDGEMENTS

The SAMI Galaxy Survey is based on observation made at the Anglo-Australian Telescope (AAT). The Sydney–AAO Multiobject Integral field spectrograph (SAMI) was developed jointly by the University of Sydney and the Australian Astronomical Observatory. The SAMI input catalogue is based on data taken from the Sloan Digital Sky Survey, the GAMA Survey and the VST ATLAS Survey. The SAMI Galaxy Survey is funded by the Australian Research Council Centre of Excellence for All-sky Astrophysics (CAASTRO, CE1100010200) and other participating institutions. The SAMI Galaxy Survey website is http://sami-survey.org. We thank the dedicated staff at the AAT whose support in interfacing SAMI with AAOmega is invaluable.

MSO and JTA acknowledge the funding support from the Australian Research Council through a Super Science Fellowship (ARC FS110200023 and FS110200013). SMC acknowledges the support of an ARC future fellowship (FT100100457). ISK is the recipient of a John Stocker Postdoctoral Fellowship from the Science and Industry Endowment Fund (Australia). LC acknowledges support under the Australian Research Councils Discovery Projects funding scheme (DP130100664). CJW acknowledges support through the Marie Curie Career Integration Grant 303912.

This research made use of ASTROPY, a community-developed core PYTHON package for Astronomy (Astropy Collaboration, 2013).

REFERENCES

Abazajian K. N. et al., 2009, ApJS, 182, 543
Allen J. T. et al., 2014, Astrophysics Source Code Library, ascl:1407.006
Allen J. T. et al., 2015, MNRAS, 446, 1567
Astropy Collaboration et al., 2013, A&A, 558, A33
Bacon R. et al., 2001, MNRAS, 326, 23
Bland J., Tully R. B., 1989, AJ, 98, 723
Bland-Hawthorn J. et al., 2011, Opt. Express, 19, 2649
Blanton M. R., Moustakas J., 2009, ARA&A, 47, 159
Brough S. et al., 2013, MNRAS, 435, 2903
Bryant J. J., O'Byrne J. W., Bland-Hawthorn J., Leon-Saval S. G., 2011, MNRAS, 415, 2173
Bryant J. J. et al., 2012, Proc. SPIE, 8446, 84460X
Bryant J. J., Bland-Hawthorn J., Fogarty L. M. R., Lawrence J. S., Croom S. M., 2014a, MNRAS, 438, 869
Bryant J. J. et al., 2014b, preprint (arXiv:1407.7335)
Cappellari M. et al., 2011, MNRAS, 413, 813
Colless M. et al., 2001, MNRAS, 328, 1039
Croom S. M. et al., 2012, MNRAS, 421, 872
Drinkwater M. J., Jurek R. J., Blake C., Woods D., Pimbblet K. A., Glazebrook K., Sharp R., Pracy M. B., 2010, MNRAS, 401, 1429
Driver S. P. et al., 2006, MNRAS, 368, 414
Driver S. P. et al., 2011, MNRAS, 413, 971
Eisenstein D. J. et al., 2005, ApJ, 633, 560
Filippenko A. V., 1982, PASP, 94, 715
Fogarty L. M. R. et al., 2012, ApJ, 761, 169
Fogarty L. M. R. et al., 2014, MNRAS, 443, 485
Fruchter A. S., Hook R. N., 2002, PASP, 114, 144
Ho I.-T. et al., 2014, MNRAS, 444, 3894
Hopkins A. M. et al., 2013, MNRAS, 430, 2047
Huchra J. P., Vogeley M. S., Geller M. J., 1999, ApJS, 121, 287
Husemann B., Kamann S., Sandin C., Sánchez S. F., García-Benito R., Mast D., 2012, A&A, 545, 137
Jones D. H. et al., 2009, MNRAS, 399, 683
Koekemoer A. M. et al., 2011, ApJS, 197, 36
Lawrence J. et al., 2012, Proc. SPIE, 8446E, 53
Moffat A. F. J., 1969, A&A, 3, 455
Rich J. A., Torrey P., Kewley L. J., Dopita M. A., Rupke D. S. N., 2012, ApJ, 753, 5
Richards S. N. et al., 2014, MNRAS, 445, 1104
Roth M. M. et al., 2005, PASP, 117, 620
Sánchez S. F. et al., 2012, A&A, 538, A8
Saunders W. et al., 2004, Proc. SPIE, 5492, 389

Sharp R., Birchall M. N., 2010, PASA, 27, 91
Sharp R. G., Bland-Hawthorn J., 2010, ApJ, 711, 818
Sharp R., Parkinson H., 2010, MNRAS, 408, 2495
Sharp R. et al., 2006, Proc. SPIE, 6269E, 14
Sharp R., Brough S., Cannon R. D., 2013, MNRAS, 428, 447
van Dokkum P. G., 2001, PASP, 113, 1420
Veilleux S., Shopbell P. L., Rupke D. S., Bland-Hawthorn J., Cecil G., 2003, AJ, 126, 2185
Welikala N., Connolly A. J., Hopkins A. M., Scranton R., Conti A., 2008, ApJ, 677, 970
York D. G. et al., 2014, AJ, 443, 485

APPENDIX A: CODE EXAMPLE FOR COVARIANCE RECOVERY

Presented below is a pseudo-code example of the covariance recovery process. While the spectral data cube is not required in the reconstitution of the covariance matrix, four other components are: the spectral variance cube, the weight map, the compressed covariance cube, and the `.FITS` header from the covariance cube extension. The procedure for regeneration of the covariance data is as follows.

(i) The covariance cube `.FITS` header is checked for the keyword `COVARMOD = Optimal`.

(ii) A full size array is created to store the reconstructed covariance.

(iii) The number of slices used to store the compressed covariance data is recovered from the covariance cube `.FITS` header via the `COVAR_N` keyword.

(iv) The indices for the wavelength slices at which the covariance map has been stored in the compressed form are extracted from the covariance cube FITS header via the `COVARLOCn` keywords – where `n` is the index of the slice in the reduced array.

(v) Running through each wavelength plane of the full size covariance array in turn, the array is populated with the corresponding slice from the compressed covariance array starting from element `0` of the compressed array, and moving to the next element of that array at the wavelength sliced indicated by `COVARLOCn`.

(vi) With the full size covariance array is now populated with the normalized data from the compressed array, one now loops over each wavelength slice in turn and multiplies the normalized mapping by the variance value, with the weight map normalization removed, for each output spaxel.

```
# Python based pseudo-code to regenerate covariance array.
# Input are:
#                   The compressed covariance array
#                   Its FITS header
#                   The full variance array
#                   The full weight map
# The full reconstructed covariance array is returned.

def reconstruct_covariance(var_array,covar_array_red,weight_array,covar_header):
    # Reconstruct the full covariance array from the reduced covariance
    # information stored in a standard cube

    if covar_header['COVARMOD'] != 'optimal':
        raise Exception('This cube does not contain covariance information in the optimal format')

    # Create an empty full covariance cube
    # This has dimensions of [Wavelength, Covariance scale, Covariance scale, Cube size, Cube size]
    covar_array_full = np.zeros([2048,5,5,50,50])

    # Populate the full covariance cube with covariance maps from reduced array
    n_covar = covar_header['COVAR_N']
    for i in range(n_covar):
        slice = covar_header['COVARLOC'+str(i+1)]
        covar_array_full[slice,:,:,:,:] = covar_array_red[i,:,:,:,:]

     # Fill values between calculated covariance map slices with
     # the last calculated value
    lowest_point = np.min(np.where((covar_array_full[1:,2,2,25,25] != 0.0) &
                                   (np.isfinite(covar_array_full[1:,2,2,25,25])))[0]) + 1
    for i in range(2048):
        if np.sum(np.abs(covar_array_full[i,:,:,:,:])) == 0:
            if i < lowest_point:
                covar_array_full[i,:,:,:,:] = covar_array_full[lowest_point,:,:,:,:]
            else:
                covar_array_full[i,:,:,:,:] = covar_array_full[i-1,:,:,:,:]

    # For each wavelength slice at each spaxel,
    # scale the normalised covariance maps by the appropriate variance removing the
    # after removing the relative exposure time weight map normalisation.
    for i in range(2048):
        for x in range(50):
            for y in range(50):
        covar_array_full[i,:,:,x,y] = covar_array_full[i,:,:,x,y] * var_array[i, x,y] * (weight_array[i,x,y]**2)

    return covar_array_full
```

This paper has been typeset from a TEX/LATEX file prepared by the author.