

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:

Metha, B;Trenti, M;Battisti, A;Chu, T

Title:

A geostatistical analysis of multiscale metallicity variations in galaxies -III. Spatial resolution and data quality limits

Date:

2024-03-01

Citation:

Metha, B., Trenti, M., Battisti, A. & Chu, T. (2024). A geostatistical analysis of multiscale metallicity variations in galaxies -III. Spatial resolution and data quality limits. Monthly Notices of the Royal Astronomical Society, 529 (1), pp.104-128. <https://doi.org/10.1093/mnras/stae491>.



Persistent Link:

<https://hdl.handle.net/11343/351750>

License:

[cc-by](#)

A geostatistical analysis of multiscale metallicity variations in galaxies – III. Spatial resolution and data quality limits

Benjamin Metha ^{1,2,3}★ Michele Trenti,^{1,2} Andrew Battisti ^{2,4} and Tingjin Chu⁵

¹*School of Physics, The University of Melbourne, VIC 3010, Australia*

²*Australian Research Council Centre of Excellence for All-Sky Astrophysics in 3-Dimensions, Melbourne, VIC 3000, Australia*

³*Department of Physics and Astronomy, University of California, Los Angeles, 430 Portola Plaza, CA 90095, USA*

⁴*Research School of Astronomy and Astrophysics, Australian National University, Cotter Road, Weston Creek, ACT 2611, Australia*

⁵*School of Mathematics and Statistics, The University of Melbourne, VIC 3010, Australia*

Accepted 2024 February 13. Received 2024 January 30; in original form 2023 October 2

ABSTRACT

Geostatistical methods are powerful tools for understanding the spatial structure of the metallicity distribution of galaxies, and enable construction of accurate predictive models of the 2D metallicity distribution. However, so far these methods have only been applied to very high spatial resolution metallicity maps, leaving it uncertain if they will work on lower quality data. In this study, we apply geostatistical techniques to high-resolution spectroscopic maps of three local galaxies convolved to eight different spatial resolutions ranging from ~ 40 pc to ~ 1 kpc per pixel. We fit a geostatistical model to the data at all resolutions, and find that for metallicity maps where small-scale structure is visible by eye (with $\gtrsim 10$ resolution elements per R_e), all parameters, including the metallicity correlation scale, can be recovered accurately. At all resolutions tested, we find that point metallicity predictions from such a geostatistical model outperform a circularly symmetric metallicity gradient model. We also explore dependence on the number of data points, and find that $N \gtrsim 100$ spatially resolved metallicity values are sufficient to train a geostatistical model that yields more accurate metallicity predictions than a radial gradient model. Finally, we investigate the potential detrimental effects of having spaxels smaller than an individual H II region by repeating our analysis with metallicities integrated over H II regions. We see that spaxel-based measurements have more noise, as expected, but the underlying spatial metallicity distribution can be recovered regardless of whether spaxels or integrated regions are used.

Key words: methods: statistical – software: data analysis – ISM: abundances – ISM: structure – galaxies: abundances – galaxies: ISM.

1 INTRODUCTION

The distribution of metals (chemical elements heavier than helium) throughout a galaxy is sensitive to many details of their formation and evolution that are not yet fully understood. It is well established that metals are released into the interstellar medium (ISM) of galaxies by supernovae (Andrews et al. 2020), but we do not know how efficiently they can mix into their local environment (Li et al. 2021). We know that a fraction of the metals released in these supernovae ejecta escape the galaxy via metal-enriched outflows (e.g. Heckman, Armus & Miley 1990; Reichardt Chu et al. 2022), but we do not know how large this escape fraction is, nor how much of the ejected material falls back on to the galaxy later in a galactic fountain (Christensen et al. 2018). We know that galaxies accrete metal-poor gas to fuel star formation (Larson 1972; Trapp et al. 2022), but there is still some debate as to whether this happens preferentially on the central regions or the outskirts of galactic discs (Schönrich & McMillan 2017). Detailed analysis of the internal chemical structure of galaxies can shed light on the relative strengths and time-scales of all of these

processes, making it a key element for addressing many theoretical challenges in understanding galaxy formation and evolution (Naab & Ostriker 2017).

Traditionally, the internal metallicity structure of galaxies has been understood through the lens of a radial linear gradient model (Searle 1971). Through a large collection of observational campaigns (e.g. Vila-Costas & Edmunds 1992; Zaritsky, Kennicutt & Huchra 1994; Croom et al. 2012; Sánchez et al. 2012; Berg et al. 2013; Bryant et al. 2015; Bundy et al. 2015; Ho et al. 2015; Belfiore et al. 2017, to name a few), it has been established that most spiral galaxies in the local universe follow a negative metallicity gradient, with more metal enrichment in their central regions than in the outskirts of their discs. This can be explained through an ‘inside-out’ galaxy formation model, in which star formation begins first in the central regions of a galaxy, allowing those regions to be more metal-enriched through successive generations of star formation than the outskirts, where star formation has only recently begun (e.g. Boissier & Prantzos 1999).

Recent technological advancements in high resolution integral field spectroscopy (IFS) has led to acquisition of detailed metallicity maps for galaxies in the local universe, with resolutions finer than ~ 100 parsec (Erroz-Ferrer et al. 2019; López-Cobá et al. 2020; Emsellem et al. 2022). Using data from MUSE, and pseudo-IFS

* E-mail: methab@student.unimelb.edu.au

data obtained through the TYPHOON survey (Seibert et al., in preparation), large azimuthal variations in metallicity have been seen to exist in many local spiral galaxies in addition to a metallicity gradient (e.g. Sánchez-Menguiano et al. 2016, 2018; Ho et al. 2017, 2018, 2019; Kreckel et al. 2019). This data has also been used to show that small-scale variations in metallicity exist in addition to a radial trend, and to analyse the spatial scale of these variations, using a variety of statistical techniques (Kreckel et al. 2020; Li et al. 2021, 2022; Williams et al. 2022).

In the first two papers of this series (Metha, Trenti & Chu 2021, hereafter M21; and Metha et al. 2022, hereafter M22), we have explored how tools and techniques from geostatistics, the mathematical study of stochastic processes that occur over a continuous spatial domain, can be applied to analyse data products from high-resolution studies of metallicity variations within star-forming galaxies in the local universe. In M21, seven galaxies from the PHANGS-MUSE survey were analysed using a geostatistical approach to recover the statistical properties of their small-scale metallicity fluctuations, allowing predictions of an analytical model to be tested. In M22, geostatistical models were fit to high-resolution metallicity maps of eight large spiral galaxies observed as a part of the TYPHOON survey, producing more accurate predictive models of the metallicity distribution throughout these galaxies that can be applied to regions obscured by diffuse ionized gas (DIG) contamination where the metallicity cannot be directly observed.

Despite these successes, a wealth of open questions remain. What are the main astrophysical sources responsible for these small-scale metallicity variations? Are they present in all galaxies, or only in large star-forming spirals in the local universe? Is the size of these small-scale variations set by the size (R_e ; effective radius) of the galaxies? Do these small-scale metallicity variations also exist in passive galaxies, or dwarf galaxies? Are star-forming galaxies with strong spiral arms expected to be more or less well mixed than those without (i.e. grand design versus flocculent)? Are these small-scale variations stronger or weaker at cosmic noon, where galaxies are more clumpy and irregular and often show flat or inverted metallicity gradients? Finally, does the 2D chemical structure of these galaxies match what is predicted by hydrodynamical simulations such as IllustrisTNG (Pillepich et al. 2018) or FIRE (Wetzel et al. 2023)?

To answer these questions and others, it is important to extend these spatial statistical analyses to a larger sample of real and simulated galaxies, with a variety of sizes, morphologies, star formation rates, and redshifts. A large wealth of legacy IFS data exists (Sánchez 2020) for which geostatistical analysis could potentially be applied. However, the application of geostatistical techniques to understand the ISM of galaxies has at present only been explored for high-resolution IFS (or IFS-like) data products.

Resolution has been identified as the most important figure of merit in the design of an IFS survey, particularly when the goal is to capture detailed information about sharp structures that vary over small scales (Mast et al. 2014). Therefore, it is important to understand how a geostatistical analysis performs on data at a range of resolutions. This will provide useful information for the design of future surveys and numerical simulations if their aims include detailed analysis of the small-scale processes that determine ISM metallicity.

Similar studies have been done in the past exploring resolution limits in IFS surveys in the context of fitting metallicity gradients to galaxies. By rebinning IFS data and simulating observations with poor angular resolution, Yuan, Kewley & Rich (2013) found that metallicity gradients appear flatter when observed with low-resolution IFS (FWHM \gtrsim 340 pc for a system at $z = 1.49$). These

effects were further investigated and quantified by Acharyya et al. (2020) using simulated high-resolution IFS datacubes. They found that input metallicity gradients are generally recovered accurately when the FWHM of the instrumental PSF is smaller than $\sim 0.25R_e$.

In this work, we investigate the impact of spatial resolution by following the methodology of Yuan, Kewley & Rich (2013) and Mast et al. (2014). Specifically, we rebin high-resolution IFS data of three galaxies observed by the TYPHOON survey to increasingly coarse resolutions, in order to assess how the results of geostatistical analysis are sensitive to resolution. Further, we explore how the quality of model fits depends on the number of metallicity data points. Finally, we provide a first investigation into the quantitative effects of aperture bias by comparing models fit using unbinned spaxels associated with H II regions to models fit using integrated H II region data.

The structure of this paper is as follows. We discuss the data and how it is processed to produce low-resolution maps in Section 2. In Section 3, we discuss our metallicity diagnostics, our methods for finding H II-dominated spaxels and H II regions, and the geostatistical model and fitting methods that we use. We present our results for NGC 5236 in detail in Section 4. We show how the ability of our geostatistical pipeline to precisely and accurately recover details on the multiscale metallicity structure of these galaxies is affected by the number of available data points in Section 4.1, and by resolution in Section 4.2. In Section 4.3, we explore the trade-off between using unbinned H II spaxels, which may be affected by aperture bias, and binned H II regions, which are fewer and necessarily at a lower spatial resolution than individual spaxels. To ensure our findings are not simply peculiarities of one galaxy, we repeat our analysis for two other spiral galaxies in Section 5. We discuss overall results in Section 6, and summarize the main conclusions in Section 7, including general advice on when and how these geostatistical methods can be effectively applied to astronomical datacubes. Finally, we present our analysis based on alternate metallicity diagnostics in Appendix A, a primer on some geostatistical techniques and concepts introduced earlier in this series in Appendix B, and a geostatistical explanation for how the variance of the metallicity measured in an IFS datacube depends on its resolution in Appendix C.

2 DATA

We select NGC 5236 (M83) as our fiducial galaxy in this study, using high-resolution spectroscopic data from the TYPHOON survey (Poetrodjojo et al. 2019). To ensure these results are valid for a range of galaxies, we also include data from two other galaxies in the TYPHOON survey, NGC 6744 and NGC 7793 (Seibert et al., in preparation). We summarize the properties of these galaxies in Table 1.

Datacubes for all three of these galaxies were constructed using a Progressive Integral Step Method (PrISM) using the 2.5-m du Pont telescope located at the Las Campanas Observatory in Chile (Grasha et al. 2022; Seibert et al. in preparation). A series of adjacent exposures were taken with a long slit of width $1''.65$, which were rebinned to pseudo-IFS datacubes with a native angular resolution of $1''.65$. TYPHOON observations were only obtained when the seeing was smaller than the width of the long slit. We note that Las Campanas Observatory has a median seeing of $0''.6$ (Persson, Carr & Jacobs 1990).

For NGC 5236 at a distance of 4.89 Mpc (Leroy et al. 2019), an angular size of $1''.65$ corresponds to a physical pixel size of 39 pc. Each datacube was rebinned to several additional resolutions by binning blocks of $f \times f$ adjacent spaxels together. We label each map

Table 1. Properties of the three galaxies explored in this paper. RA and Dec. values are taken from the NASA/IPAC Extragalactic Database. Morphological classification and position angle data are taken from HyperLEDA. Inclinations are calculated based on the axis ratios reported by HyperLEDA, assuming a disc scale height of $q_z = 0.2$ for all galaxies, following Battisti, Calzetti & Chary (2017). Distances, stellar mass estimates, and SFR estimates are taken from Leroy et al. (2019). Half-light radii, also known as effective radii (R_e) are taken from the K_S band effective radii measured from the 2MASS large galaxy atlas (Jarrett et al. 2003). B -band isophotal radii (R_{25}) are taken from de Vaucouleurs et al. (1991). Both radii have been converted to physical units based on the adopted distance values.

Name	NGC 5236	NGC 6744	NGC 7793
RA	13:39:55.96	19:09:46.10	23:57:49.83
Dec.	-29:51:55.5	-63:51:27.1	-32:35:27.7
Distance (Mpc)	4.9	11.6	3.6
R_e (kpc)	3.3	7.2	1.8
R_{25} (kpc)	9.18	33.7	4.89
Position angle	54.0°	13.7°	94.4°
Inclination	12.5°	53.5°	55.6°
Morphology	Sc	Sbc	Scd
Stellar mass ($\log_{10} M/M_\odot$)	10.4	10.9	9.2
Star formation rate ($M_\odot \text{ yr}^{-1}$)	4.0	4.0	0.25

according to its *bin factor* f . In this study, bin factors of 1, 2, 4, 6, 8, 10, 12, and 24 are tested, corresponding to physical resolutions from 39 to 940 pc for NGC 5236. This selection of resolutions covers the range of physical resolutions from MUSE observations of local galaxies, to high-resolution zoom-in cosmological simulations such as AURIGA (125 pc, Grand et al. 2017), to IFS observations of gravitationally lensed galaxies with the NIRISS instrument of *JWST* (~ 200 pc; Wang et al. 2022), up to large IFS surveys of galaxies with coarse spatial resolution such as CALIFA (~ 800 pc; Espinosa-Ponce et al. 2020). Therefore, the range of spatial resolution considered allows us to explore for which, if any, of these kinds of data products and observational surveys a geostatistical analysis is feasible.

To illustrate what features of the galaxy are visible, in Fig. 1 we show $H\alpha$ images for NGC 5236 at each resolution explored. At the native resolution, individual H II regions can be seen as ‘bubbles’ along the spiral arms of this galaxy, and regions of low $H\alpha$ emission in the inter-arm regions are visibly resolvable. As f is increased, structural details of this galaxy are lost. At $f = 4$ (with a spatial resolution of ~ 150 pc), individual H II regions begin to blur together, but variations can still be seen along spiral arms. These small-scale variations become difficult to see at $f = 8$, but the azimuthally varying spiral structure is still visible. At the coarsest resolution ($f = 24$), no azimuthal structure is apparent, and a radial trend of decreasing $H\alpha$ brightness is all that can be seen.

3 METHODS

For each rebinned datacube, LZIFU (Ho et al. 2016) was used to find the intensities of 13 strong emission lines, including $H\alpha$, $H\beta$, $[\text{O III}] \lambda 5007$, $[\text{S II}] \lambda \lambda 6716, 31$, and $[\text{N II}] \lambda 6583$. The strength of all emission lines is corrected for dust attenuation assuming Case B recombination, fixing the Balmer decrement to be $H\alpha/H\beta = 2.86$, using the extinction curve of Cardelli, Clayton & Mathis (1989), assuming $R_V = 3.1$ to match the Milky Way value.

3.1 H II spaxels selection

Strong emission line ratio based metallicity diagnostics are only valid for H II regions – regions recently ionized by hot, bright,

young O/B stars. We ensure that all spaxels used in this analysis are predominately ionized by H II region emission at each resolution by imposing data quality cuts designed to isolate H II regions from diffuse ionized gas (DIG), a significant source of contamination for metallicity studies (e.g. Reynolds 1990; Kumari et al. 2019; Poetrodjojo et al. 2019, M22).

We use two BPT diagram-based cuts (Baldwin, Phillips & Terlevich 1981), restricting our analysis to all spaxels that lie below both the Kewley et al. (2001) demarcation line on the $[\text{S II}]$ -based BPT diagram and the Kauffmann et al. (2003) demarcation line on the $[\text{N II}]$ -based BPT diagram. To ensure these cuts are robust, only spaxels with measurements of $H\alpha$, $H\beta$, $[\text{O III}] \lambda 5007$, $[\text{S II}] \lambda \lambda 6716, 31$, and $[\text{N II}] \lambda 6583$ fluxes with $S/N > 3$ are considered. Following Zhang et al. (2017), we additionally impose a surface brightness cut, including only spaxels with a $H\alpha$ surface brightness greater than $10^{39} \text{ erg s}^{-1} \text{ kpc}^{-2}$. After these cuts, for NGC 5236, 7488 spaxels are classified as being predominately ionized by H II emission at native resolution. We hereafter refer to such spaxels as H II spaxels. As f is increased to simulate coarser resolution observations, the number of H II spaxels recovered by this methodology decreases. We plot the number of H II spaxels available for analysis as a function of the resolution of the binned datacubes for all three galaxies analysed in this work in Fig. 2.

From observations of H II regions in the Milky Way, it has been established that the spatial extent of a H II region can range from tens of parsecs to kpc scales (Cosens et al. 2018), with typical sizes ranging from 80 to 800 pc (Espinosa-Ponce et al. 2020). Therefore, the high-resolution metallicity maps examined in this work contain H II spaxels that are significantly smaller than an individual H II region. This may introduce bias into measurements based on individual high-resolution spaxels, as emission line ratios are not constant throughout a H II region (e.g. Ferland et al. 2013), and strong emission line based metallicity diagnostics are calibrated to work on integrated H II regions (e.g. Mannucci et al. 2021).

To assess whether this bias is significant, we additionally extract metallicities for integrated H II regions detected using the publically available code PYHIIEXTRACTOR (Lugo-Aranda et al. 2022). We choose this program over other H II region detection software because it does not require a license, it is actively maintained, and it was designed for optimal performance on high-resolution IFS data. Based on the $H\alpha$ intensity map and its uncertainty, positions and radii of hundreds of H II region candidates are extracted from each galaxy at native resolution. The strengths of the emission lines for each candidate H II region are then computed by summing all (extinction-corrected) fluxes from spaxels contained within each region.

One limitation of PYHIIEXTRACTOR for our application is that the same spaxel can be contained within two (or more) different overlapping extracted regions. In these cases, we assign spaxels to the H II region whose centre is closest to them. This step is necessary to prevent spurious correlations arising between neighbouring H II regions that are not physical, but merely statistical artefacts from double-counting.

We plot all H II region candidates on the two BPT diagrams, discarding any regions lying above the Kewley et al. (2001) demarcation line on the $[\text{S II}]$ -based BPT diagram and the Kauffmann et al. (2003) demarcation line on the $[\text{N II}]$ -based BPT diagram, leaving a final sample of H II regions for each galaxy. We note that this processing step decreases both the number of data points available for a metallicity model fit and the effective resolution of the data, but may be appropriate in some cases to mitigate against aperture bias. We explore how including this processing step affects a geostatistical model fitting procedure in Section 4.3.

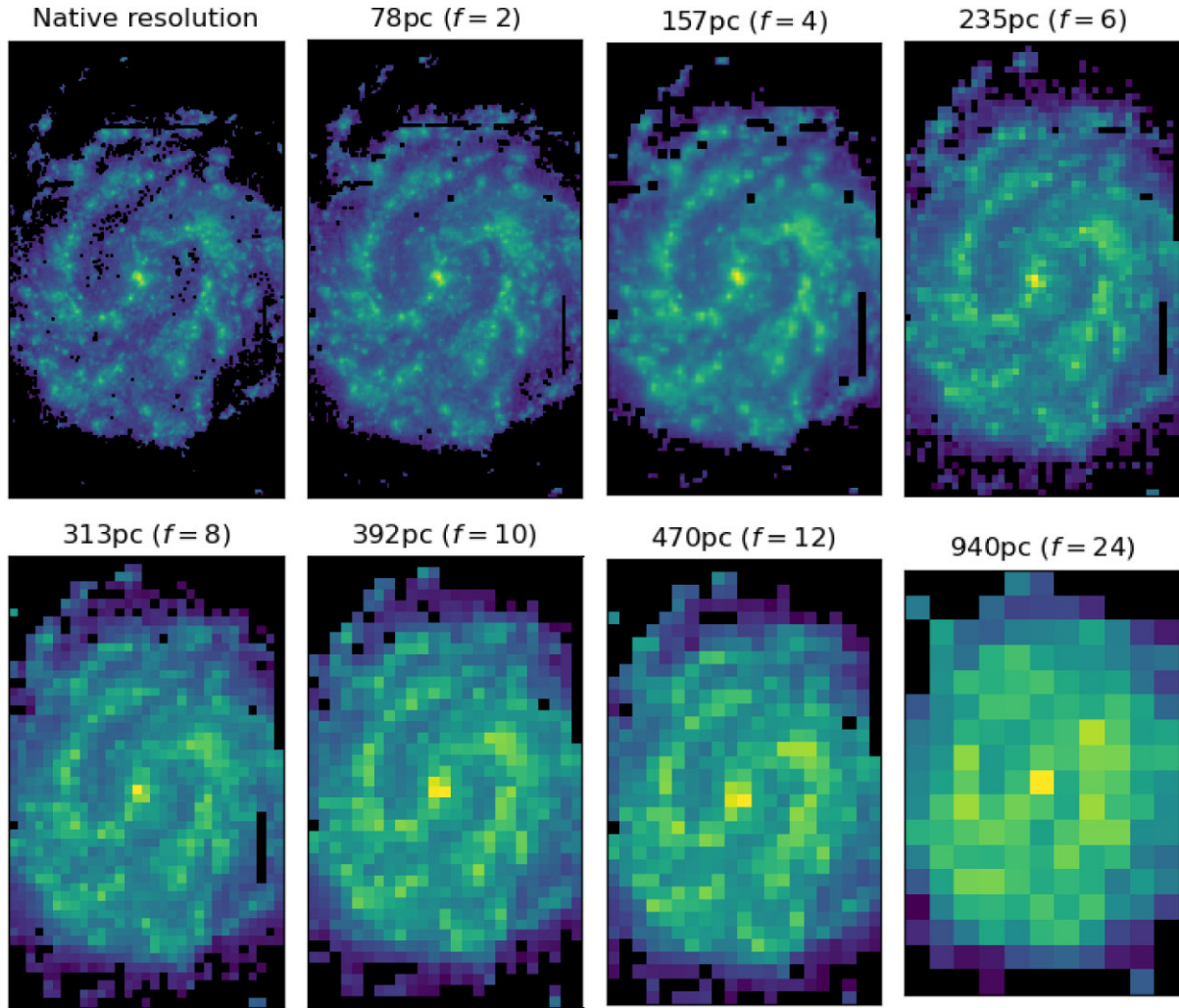


Figure 1. Surface brightness of H α for NGC 5236 at its native resolution (top left), and for spatially binned datacubes containing the total emission from $f \times f$ spaxels in each pixel. For the very high resolution maps ($f = 1 - 2$), individual H II regions may be resolved clearly. On larger scales ($f = 4 - 6$), bright regions are blurred together, but small features are still visible. For intermediate resolution maps ($f = 8 - 12$), small-scale features become harder to see, but large-scale features such as spiral arms are still visible. With low-resolution IFU observations ($f = 24$), all features apart from a radial trend in H α brightness are lost.

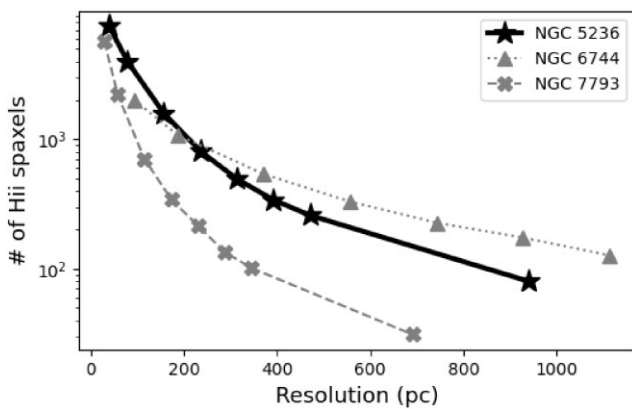


Figure 2. Number of H II spaxels as a function of resolution. For each galaxy, as the resolution of the data is decreased, the number of spaxels classified as H II dominated also decreases.

3.2 Metallicity diagnostics

To ensure that our results are robust, we investigate using four different metallicity diagnostics – the $N_2S_2H\alpha$ diagnostic with the calibration of Dopita et al. (2016), the $Scal$ diagnostic of Pilyugin & Grebel (2016), and the O_3N_2 and RS_{32} diagnostics of Curti et al. (2020). When using each diagnostic, we discard any H II spaxels (or regions) that have $S/N < 3$ in any of the lines used. This final data quality cut excludes < 1 per cent of the data. For all of the diagnostics we explore, we see similar trends for how the geostatistical methods tested are affected by resolution and data quality. We choose $RS_{32} = \log([\text{O III}] \lambda 5007 / H\beta + [\text{S II}] \lambda \lambda 6717, 6731 / H\alpha)$ as our fiducial diagnostic, presenting results from this diagnostic in the main text. This diagnostic is calibrated using direct electron temperature-based metallicity estimates of stacked spectra of SDSS galaxies with a 4th order polynomial fit for Z as a function of RS_{32} . We refer the reader to Appendix A for details on the other metallicity diagnostics investigated in this work, as well as for a presentation of the main results for NGC 5236 when these alternative metallicity diagnostics are used.

3.3 Geostatistical model fitting

In this Section, we introduce the hierarchical geostatistical model that we fit to our data, and the parameter determination procedure. The model is identical to the one introduced in M22. However, the fitting procedure has been updated and improved. For a summary on key background on other tools from geostatistics that are used in this work and have been introduced in the earlier papers of this series, we refer the reader to Appendix B.

For each data set that we test, we consider the observed metallicity Z_{obs} at each location \vec{x} to be a combination of the true underlying metallicity at that point ($Z(\vec{x})$), plus some uncorrelated, stationary, zero-mean observational noise $\epsilon(\vec{x})$:

$$Z_{\text{obs}}(\vec{x}) = Z(\vec{x}) + \epsilon(\vec{x}). \quad (1)$$

We further separate $Z(\vec{x})$ into two components: the process mean $\mu(\vec{x})$, which captures the large-scale metallicity trends of the galaxy, plus a random component $\eta(\vec{x})$ with zero-mean, which represents real, small-scale metallicity fluctuations around this mean:

$$Z(\vec{x}) = \mu(\vec{x}) + \eta(\vec{x}). \quad (2)$$

As in M21, M22, $\mu(\vec{x})$ is modelled as being linearly dependant on the distance between \vec{x} and the centre of the galaxy. To fully specify this model, two parameters must be fit: ∇Z , the radial metallicity gradient, and the characteristic metallicity Z_{char} at a user-defined radius r_{char} :

$$\mu(\vec{x}) = Z_{\text{char}} + \nabla Z \cdot (r(\vec{x}) - r_{\text{char}}). \quad (3)$$

In M21, M22, and in most studies when metallicity gradients are fit, r_{char} is chosen to be zero. Z_{char} therefore becomes the central metallicity of the galaxy. However, we found that such a choice of parametrization led to strong anticorrelation between Z_{char} and ∇Z , with $\rho < -0.9$ between these two parameters irrespective of the galaxy under investigation, the metallicity diagnostic used, or the resolution explored. Such a strong anticorrelation has been noted before in the literature, e.g. by Zaritsky, Kennicutt & Huchra (1994).

To mitigate this anticorrelation, and in turn improve the efficiency of convergence using Monte Carlo methods, we parametrize the linear function by fitting the characteristic metallicity at a radius of $r_{\text{char}} = 0.4R_{25}$ (see Zaritsky, Kennicutt & Huchra 1994). The metallicity at this radius has been observed to approximately match the integrated metallicity of face-on spiral galaxies using a variety of metallicity calibrations (e.g Moustakas & Kennicutt 2006; Rosales-Ortega et al. 2011). We found that such a parametrization significantly lowered the degree of correlation between Z_{char} and ∇Z , for all fits of our galaxy maps.

We parametrize the covariance of the small-scale, random metallicity fluctuations $\eta(\vec{x})$ as an exponential function, following M22. Such a functional form was chosen because in the small-scale limit, the structure of this field will mirror that of a passive scalar driven by Burger's turbulence (Lee & Gammie 2021), which is a good model for shock-induced structures (Bec & Khanin 2007; Falceta-Gonçalves et al. 2014), and may be the most suitable paradigm of turbulence for describing the supernova shock-driven mixing of passive scalars around H II regions. This function has two parameters: ϕ , which sets the distance at which random metallicity fluctuations are correlated, and σ^2 , the total variance about the process mean. For two data points \vec{x}, \vec{y} separated by a distance of h , we model the covariance between the small-scale metallicity fluctuations to be:

$$\text{Cov}(Z(\vec{x}), Z(\vec{y})) = \sigma^2 \exp\left(-\frac{h}{\phi}\right). \quad (4)$$

All four parameters of this hierarchical model ($Z_{\text{char}}, \nabla Z, \phi$ and σ^2) are fit simultaneously using a maximum-optimized method, optimized using the Python package EMCEE (Foreman-Mackey et al. 2013). We refer the reader to equation (6) of M22 for the full form of the 4D likelihood function that is maximized. To help distinguish between values of σ^2 that are close to zero, and to ensure that the value of σ^2 was always positive, $\log(\sigma^2)$ was fit instead of σ^2 , with flat priors.

For every data set explored, we estimate posterior distributions for each parameter by running EMCEE for 480 steps using 72 walkers. Exploratory work showed that the autocorrelation time, τ_{AC} , for each parameter is ~ 40 steps. Based on this, 480 steps was chosen such that the program could run for $10\tau_{AC}$ steps, after discarding $2\tau_{AC}$ steps for burn-in. This ensures there are effectively ~ 720 independent samples from the posterior distribution of each parameter, which is sufficient for estimating parameters with realistic uncertainties for this 4D problem. Each Monte Carlo chain was examined by hand to ensure that this burn-in length was long enough; for cases where it was not, an additional 40 steps were ran so that $3\tau_{AC}$ steps could be discarded.

For each run, walkers were initially started at a wide range of initial values of ϕ (ranging from 0.1 to 1.6 kpc) and σ^2 , in order to ensure solutions were not confined to local minima, and a uniform prior was enforced on ϕ to keep it bounded between 50 pc and 5 kpc. The lower limit was chosen as it is similar to the size of a single pixel at native resolution for this data set, and it is not possible to observe fluctuations that are smaller than a single pixel (Cressie 1993). For initial guesses of Z_{char} and ∇Z , we use the results of a simple weighted-least squares (WLS) fit, with an uncertainty of 0.05 dex on Z_{char} and an uncertainty of 0.01 dex/kpc on ∇Z , as central metallicities and metallicity gradients recovered for galaxies using the WLS method have been found to agree with those found with a geostatistical method to these levels (M22). After a test run, the chains of each Monte Carlo run were examined for convergence. We found that for some runs, all chains converged to low values of ϕ except for those that were started at values of ϕ that were much larger than the best-fitting value. In these cases, we re-ran an MCMC fit using a set of initial values for ϕ ranging from 0.1 kpc to 0.45 kpc. In a few cases, the upper limit on the prior on ϕ was also lowered to 3 kpc, to ensure all chains remained close to the best-fitting value, as informed by previous runs. We note that adjusting parameter ranges based on information inferred from previous fits is a standard practice in a Bayesian framework, where the aim is to achieve quicker convergence on the global minimum while efficiently using the (limited) computational resources available.

To illustrate the quality of our fits, in Fig. 3 we show the approximate posterior generated by EMCEE as a corner plot for NGC 5236, using all H II spaxels, our fiducial metallicity diagnostic of RS_{32} , and a datacube without any additional binning. From this plot, we see that the two small-scale parameters, ϕ and $\log(\sigma^2)$, are strongly correlated with each other ($\rho = 0.94$). As ϕ is increased, nearby data points are expected to become more and more correlated with each other, reducing the overall variability of the model. To match the amount of variation that is seen in the data set that is being fit, σ^2 must be increased to compensate for this effect. Therefore, models that fit the data and have a high value of ϕ must have a lower value of σ^2 , and vice versa. Such behaviour is expected, and does not impact on the results presented in the following Sections.

4 RESULTS

In this Section, we assess how the performance of the geostatistical methods explored for galactic metallicity modelling in M21 and

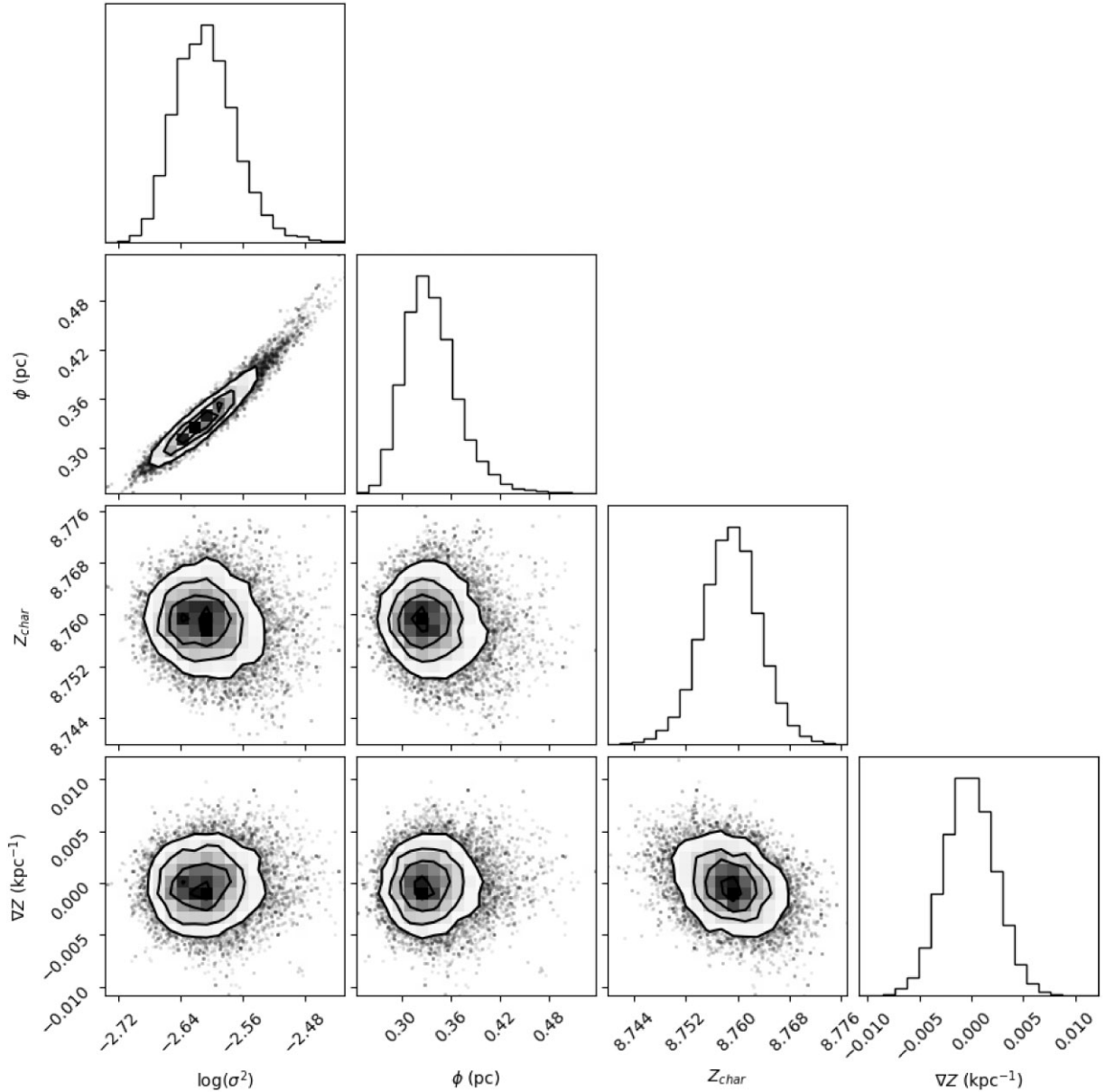


Figure 3. Corner plot, showing 1D and 2D marginalized posterior distributions for the four parameters of our geostatistical model, using all H II spaxels from our fiducial galaxy (NGC 5236), with our fiducial metallicity diagnostic (RS_{32}), at the native resolution (39 parsec per spaxel). A strong correlation can be seen between ϕ and $\log(\sigma^2)$, the two parameters that describe small-scale metallicity fluctuations within the galaxy, as raising ϕ will naturally lower the variability of metallicity within the data. Such a correlation is seen in all EMCEE fits performed in this work.

M22 depends on three facets of the datacube used: the number of data points available, the resolution of the data, and whether the fit is performed on individual H II spaxels, or over integrated H II regions. However, separating out these effects is difficult. When spaxels are binned to make coarser datacubes, the number of data points available for statistical inference and model fitting decreases (see Fig. 2). Furthermore, integrating H II spaxels into regions both decreases the number of available data points for model fitting, and the resolution of the available data points.

With this in mind, in Section 4.1, we first assess how these methods are affected the number of data points with measured metallicity values available. We test this by randomly selecting subsamples of a limited number of H II spaxels from our flagship galaxy NGC 5236 at native resolution, allowing us to change the number of available

data points for analysis without introducing any additional spatial effects.

Then, in Section 4.2, we report on how the model parameters recovered and the accuracy of metallicity predictions depend on the spatial resolution of the data by comparing results for our binned datacubes for NGC 5236 to the results found at native resolution. By comparing the results for the rebinned datacubes to the results on the native resolution data sets of Section 4.1, we can tell if there are any additional effects caused by having coarsely resolved data that would not be expected by merely changing the number of data points available.

Finally, in Section 4.3, we see if the results are the same for H II regions as they are for H II dominated spaxels, by comparing spaxels that pass the BPT-diagnostic cuts of Kewley et al. (2001) and

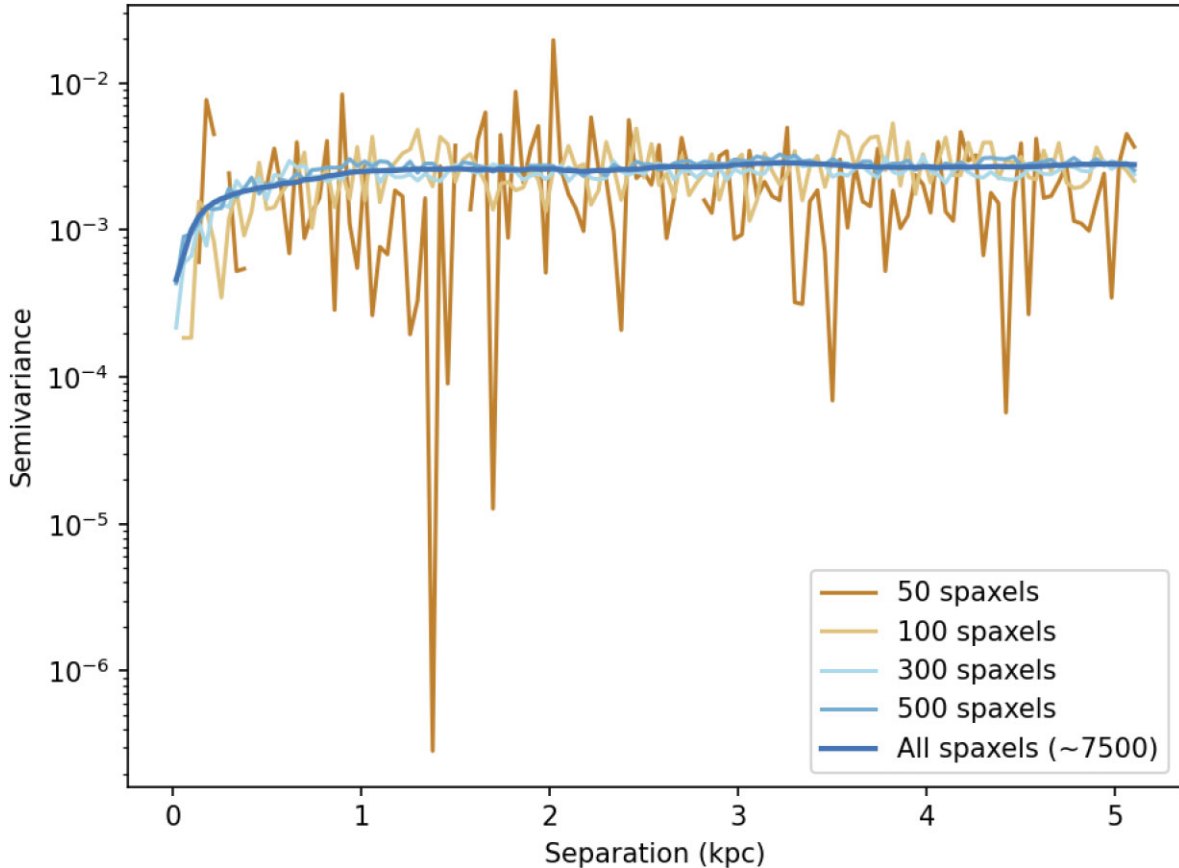


Figure 4. Recovered semivariograms for NGC 5236, when only a sub-sample of spaxels are used. In all cases, the semivariogram has the same shape – however, when the number of data points is ≤ 100 , small number statistics lead to unreliable estimates of the variance between pairs of spaxels at each separation, increasing the noise in the semivariogram, making its interpretation less clear.

Kauffmann et al. (2003) to regions found using PYHIIEXTRACTOR that also pass these cuts (for further details, see Section 3.1). We interpret the results of this experiment in light of what we have learned from Sections 4.1 and 4.2, to see if binning the data into H II regions has any impact on the output of the model that cannot be explained by the other two effects investigated in this work.

4.1 The effects of missing data

To test how geostatistical methods perform when the number of available H II spaxels is small, we generate random subsamples of 50, 100, 300, and 500 H II spaxels from NGC 5236. We compare the performance of our geostatistical methods on these subsampled datacubes to the full datacube, which has 7454 H II spaxels for which metallicities could be determined using RS₃₂.

We create a semivariogram using each of the subsamples of the native resolution datacube in Fig. 4, and plot them against the original semivariogram computed for this data. Each semivariogram shows how the variance in the metallicity between spaxels increases as the distance between spaxels is increased, after accounting for the large-scale metallicity gradient.¹ A bin size of 40 pc, chosen to match the

¹ We provide a formal mathematical definition of the semivariogram, an introduction to this method, and its relation to some other tools used in astronomy to extract the covariance structure of random fields in Appendix B1.

resolution of the data and the minimum separation between spaxels, is used for all semivariograms.

We see that all semivariograms have the same general shape, indicating that the same small-scale structure is being detected in all five data sets. However, we also see that as the number of available data points decreases, the noise in the recovered semivariogram increases. When only 50 data points are used, some separations have missing values for the semivariance, as there are not enough pairs of H II spaxels separated by approximately that distance for the variance between them to be estimated robustly. This issue could be partially overcome by creating a semivariogram with a coarser binning, allowing a more clear view of larger features, but doing so would result in less retention of information about the small-scale variability of the data. Since we are primarily investigating how many data points are necessary to recover these small-scale properties, we do not do this.

In Fig. 5, we compute estimates for all four parameters of our geostatistical model for each subsample of the native resolution data set, and investigate the convergence of these parameters. We see that even with as few as 50 data points, the properties of the large-scale variability (∇Z and Z_{char}) are recovered accurately. In fact, ∇Z is consistent between all models, and Z_{char} varies by only 0.02 dex. Similarly, all models converge to the same value for the amount of scatter around the mean trend (σ^2), although the uncertainty on this parameter increases as the number of available data points decreases.

Recovered values on the metallicity correlation scale ϕ are consistent when 300 or more data points are available for fitting.

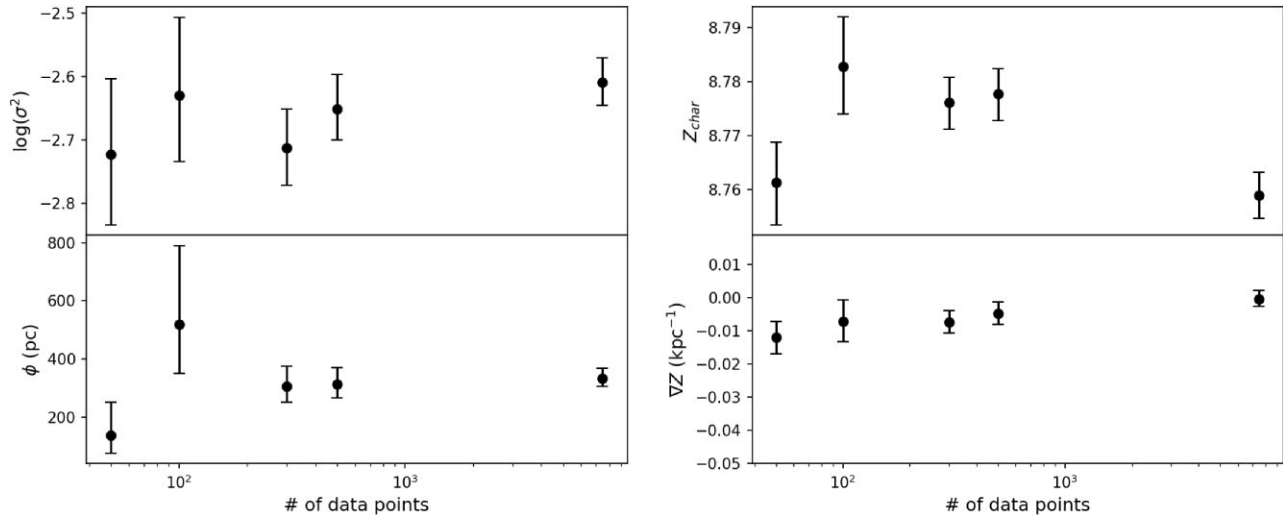


Figure 5. Best-fitting values, with uncertainty, of the metallicity correlation scale ϕ , the (log) total variance σ^2 , the characteristic metallicity Z_{char} , and the metallicity gradient ∇Z for NGC 5236 as a function of the number of spaxels at native resolution used for the fit. Generally, we find that the best-fitting parameter values are consistent with the values recovered when all spaxels are used (far right data point), albeit with large scatter when the number of available data points is low.

When only 100 data points are used, the best-fitting value of this parameter is larger, with larger uncertainties ($\phi = 518_{-166}^{+271}$ pc), but it is still consistent with the value found at native resolution ($\phi = 333_{-28}^{+34}$ pc) within the 1σ error bars. For the model trained on only 50 spatially resolved metallicity measurements, a smaller estimate of ϕ is obtained, with smaller error. By examining the posterior distribution of the Monte Carlo fit for this trial, we find that this is due to estimates clustering tightly to the lower limit of ϕ at 50 pc, indicating that no evidence for spatial correlations is seen in this limited data set.

Next, we evaluate how the goodness-of-fit of our geostatistical model is affected by having a limited number of data points. We judge the quality of our model fits over each subsampled datacube in terms of their predictive power. Following M22, we assess this using ten-fold cross-generalize. This method is a standard technique to estimate the ability of a model to generalise to unknown data points (Field, Miles & Field 2012, Chapter 7.7.2.), and is recommended as a reliable general method for model selection (Hogg & Villar 2021). For each subsampled datacube, the H II spaxels are split into ten evenly sized groups. Data from all but one group are used to fit parameters for the geostatistical model described in Section 3.3, and generate point predictions for the metallicities of the H II spaxels in the final group via the process of universal kriging, outlined in Appendix B2. This is repeated for all ten groups to produce a predicted metallicity for every spaxel.

We summarize the results of this exercise by looking at the MAD of the predicted metallicity of each spaxel to the metallicity observed for that spaxel, which we use as a metric for the model’s performance. We then normalize this value to the variance of the data by dividing by one standard deviation, σ , to make this metric easier to interpret.²

As a comparison point, we also repeat this cross-validation process for a metallicity gradient model with no small-scale component, fit for each set of data points using the standard weighted least squares (WLS) method, implemented in the Python package STATSMOD-

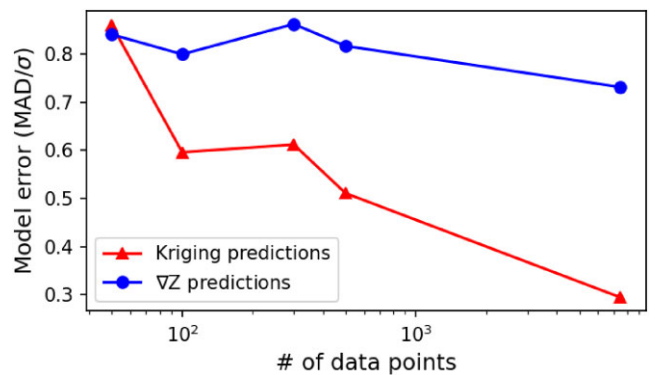


Figure 6. The normalized prediction error (median absolute deviation per standard deviation) of the predictions for metallicity computed using the cross-validation method for NGC 5236 using a simple linear gradient (blue line) and a geostatistical model (red line). As long as the number of data points $\gtrsim 100$, predictions are more accurate when a geostatistical model including small-scale metallicity variations was included.

ELS. In the language of our model, this is equivalent to fixing $\eta(\vec{x}) = 0$.

We plot MAD/ σ for the linear gradient model and the geostatistical model as a function of the number of data points available in Fig. 6. For the linear model with no small-scale component, we do not see a trend with the prediction accuracy model changing with the number of data points over the range we have chosen. This implies that 50 data points is sufficient to construct a metallicity gradient model; and using more data points in the fit will not make this model any better. On the other hand, we see the accuracy of the geostatistical model scales with the number of data points available for training. With only 50 spatially resolved metallicities, the metallicity predictions of a geostatistical model have comparable accuracy to the predictions from a metallicity-gradient model. With 100–300 data points, the geostatistical model begins to outperform the azimuthally symmetric model that only captures large-scale metallicity variations within a galaxy. With a larger sample of data points, the improvement increases dramatically.

²For reference, data from a Gaussian noise distribution around a known mean is expected to have $\text{MAD}/\sigma = \sqrt{\frac{2}{\pi}} \approx 0.8$.

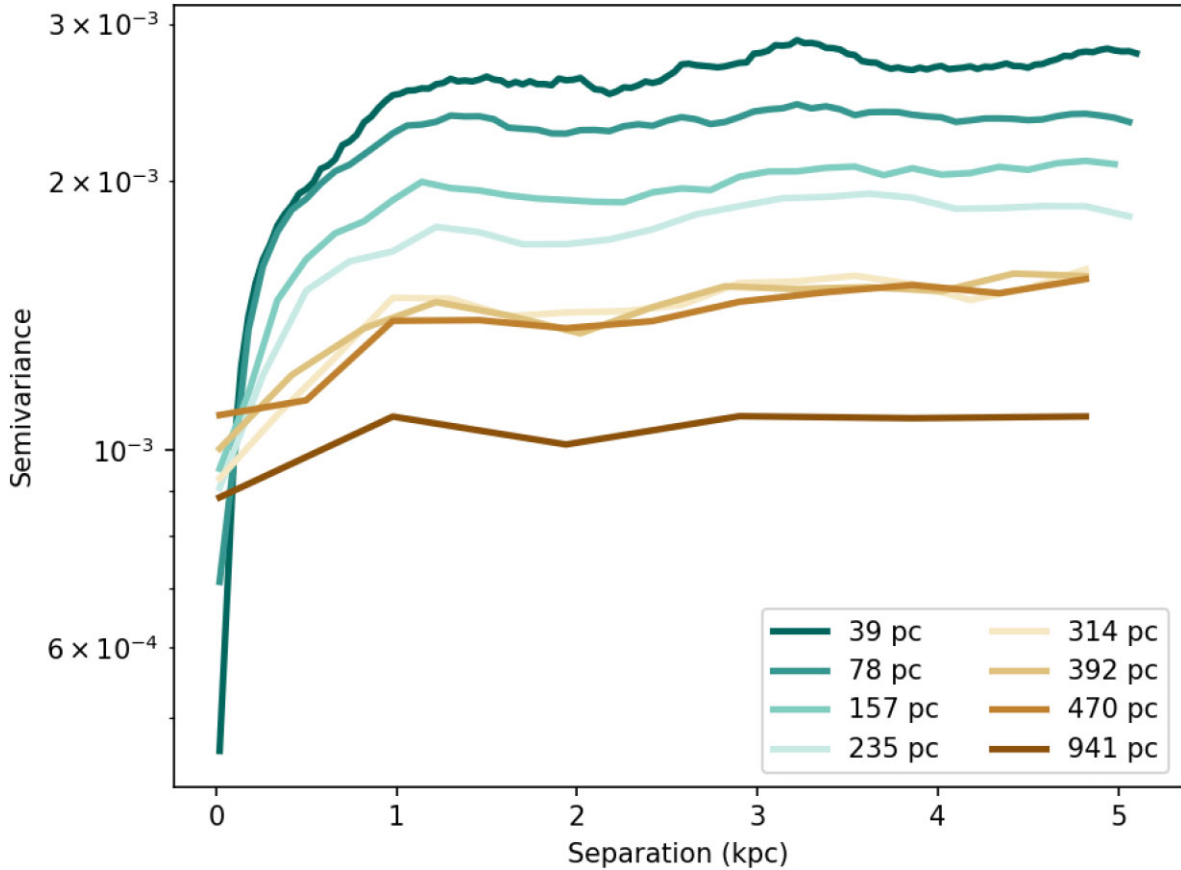


Figure 7. The semivariogram for the TYPHOON metallicity map of NGC 5236, binned to eight different spatial resolutions. As the resolution becomes coarser, the total variance in metallicity between spaxels decreases, as small-scale inhomogeneities are binned together. When the spatial resolution is finer than ~ 250 pc ($f \leq 6$), the shape of the semivariogram matches that of the native resolution map, indicating that the resolution is fine enough for small-scale metallicity fluctuations to be accurately resolved. For resolutions of 300–400 pc per pixel, the general shape of the semivariogram can still be seen, and an estimate of the scale at which metallicities become locally uncorrelated can still be made. At resolutions coarser than this, it is difficult to infer the presence of small-scale fluctuations, and accurately describe their properties.

This result is expected from geostatistical theory. Kriging works by taking advantage of the fact that nearby data points are expected to have similar properties (Tobler 1970, Appendix B2). When more data points are used, it is more likely that each data point lies within $\lesssim \phi$ parsecs of another spaxel with known metallicity, allowing the metallicity to be inferred using local galaxy properties as well as global ones. Furthermore, when more data points are used, the recovered values of the small-scale structure parameters ϕ and σ^2 are more accurate.

4.2 The effects of spatial resolution

To determine the resolution limit at which small-scale metallicity variations are visible, we construct a semivariogram for the small-scale metallicity variation using data convolved to each binning factor f . We plot these results in Fig. 7.

The most striking trend that can be seen in this series of plots is that the height of the semivariograms at large separations decreases as f is increased. This implies that the total amount of variance about a linear gradient is lower when the data is of a coarser quality. Qualitatively, this is to be expected. When the spatial resolution of IFU spectroscopy is coarser than the size of metallicity variations, different features are averaged together over each individual spaxel, decreasing the variance around a galaxy-scale metallicity model.

Quantitatively, we explore how this spatial averaging would impact the variance seen at different resolutions if the true small-scale metal variations follow an exponential correlation function with a correlation scale of $\phi = 330$ pc, as is suggested by our highest resolution data, in Appendix C. We find that the reduction of variance seen at different resolutions matches the reduction predicted with this model of spatially correlated metallicity variations very tightly.

The separation at which the semivariance stops increasing (called the *range* of the data in geostatistics literature; e.g. Cressie 1990) is not seen to increase with f for all semivariograms computed with $f \leq 6$ (corresponding to physical resolutions finer than 250 pc per pixel, plotted as blue lines in Fig. 7). Visually, when $f \leq 6$, all semivariograms show the same shape, beginning to curve downwards at a separation of ~ 0.4 kpc, before flattening out completely by ~ 1 kpc. For coarser data ($f > 6$, plotted as brown lines in Fig. 7), this shape is no longer clearly visible. At $f = 8$, flattening is still seen at ~ 1 kpc, but the shape of the semivariogram for sub-kpc data is no longer accurate. At $f > 8$, we see the apparent range of the data increase with f , as trends associated with physical small-scale variations become washed out, and the size of the pixels becomes the most important scale for setting the observed structure of the metallicity map.

To show this result in a more quantitative way, we present the best-fitting value of ϕ , the metallicity correlation scale, with uncertainty, as a function of the resolution of available data in Fig. 8. Using the

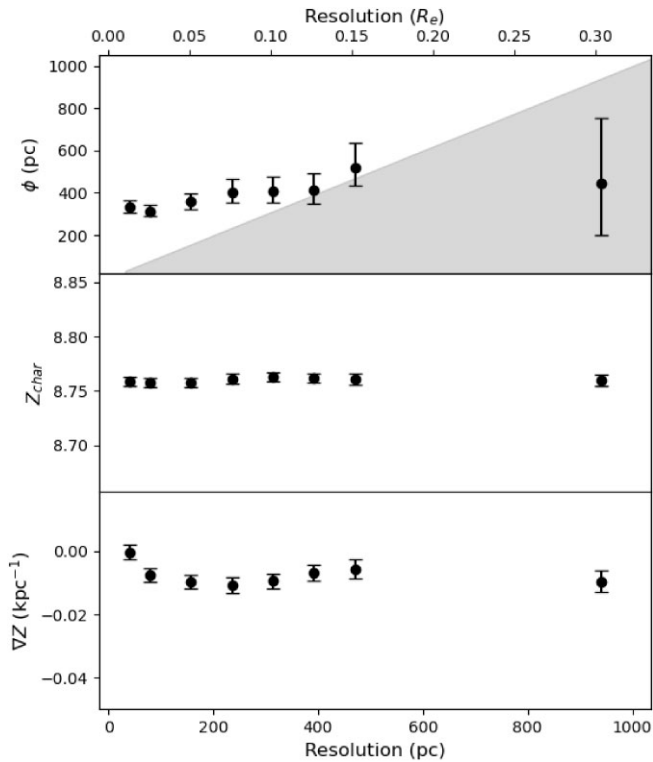


Figure 8. Best-fitting values, with uncertainty, of the metallicity correlation scale ϕ , the characteristic metallicity Z_{char} , and the metallicity gradient ∇Z for NGC 5236 as a function of the spatial resolution of metallicity maps, reported in parsec (bottom axis) and effective radii (R_e , top axis). In the top panel, the grey shaded region shows the area of parameter space where the correlation scale ϕ is smaller than the size of a pixel. We find that kpc-scale resolved metallicity surveys are able to accurately recover large-scale trends of the metallicity profile (Z_c , ∇Z), while the correlation scale can only be accurately constrained by observations with a resolution $\lesssim 400$ pc, or $0.12R_e$.

native resolution data, the most likely value of ϕ is 332^{+33}_{-28} pc. As f is increased from 1 to 10, the best-fitting value of ϕ increases slightly and has slightly larger uncertainties, but remains consistent with the value found for native resolution data. At $f=12$, the size of a single spaxel becomes 470 pc – larger than the value of ϕ recovered for the high-resolution maps. At this resolution, our model fitting procedure returns a value of $\phi = 527^{+135}_{-90}$ pc, inconsistent with the value of ϕ suggested by the higher resolution maps. At our coarsest resolution, ϕ is inferred to be 443^{+308}_{-244} pc. While this value is consistent with the ϕ recovered at the finest resolutions, the uncertainties are very large, reflecting the fact that the value of ϕ cannot be well-constrained by this data, as a single pixel is expected to be much larger than the typical size of small-scale variations.

Fig. 8 also shows that the recovered values of Z_{char} and ∇Z are relatively consistent throughout all data maps explored, excepting the results found for the datacube at native resolution, where a flatter metallicity gradient is preferred. We note that the fitted parameters for this model at native resolution are the same as those shown in Fig. 5. On its own, this result is not significant – while the values of Z_{char} and ∇Z recovered using our EMCEE-based parameter fitting pipeline do not agree with any other recovered values of Z_{char} and ∇Z at other resolutions to 1σ , we would not expect the 1σ credible intervals of eight different model fits to all overlap. However, we also find a similar result of the recovered value of ∇Z being flatter when two out of our three alternative metallicity diagnostics are used ($\text{N}_2\text{S}_2\text{H}\alpha$

and Scal , with results inconclusive for O_3N_2 ; see Appendix A), and for the galaxy NGC 7793 when the size of a spaxel is $\lesssim 100$ pc (see Section 5). While formally significant, the size of this deviation is small – smaller than the difference between metallicity gradients recovered for this galaxy when different metallicity diagnostics are used (see fig. 9 of Poetrodjojo et al. 2019).

This result is unexpected from previous studies of the effects of resolution on recovering the metallicity gradient. In fact, previous research has shown that a flattening of the metallicity gradient is expected for poorly spatially resolved data due to the effects of pixelization and beam smearing (e.g. Yuan, Kewley & Rich 2013; Acharyya et al. 2020). However, in this case, it is not a degradation of the data that causes an apparent flattening of the observed gradient, as a tendency for Z to decrease as r increases is still seen even in the native resolution datacube.

To test the validity of our statistical pipeline, we analysed the values of ∇Z recovered using a simple WLS approach without incorporating any small-scale variation. We present our results in Fig. 9. At each resolution, we plot the measured metallicity of each H II spaxel, with error bars, at its distance from the galaxy centre. We overplot the best-fitting linear relations describing the large-scale metallicity trends recovered using both our geostatistical approach, and using least squares. For all of the binned data sets ($f > 1$), the profiles recovered using WLS and our geostatistical method show good agreement. However, when $f = 1$, the geostatistical approach prefers to describe the data with a flatter metallicity gradient. This difference is significant, with $p < 5 \times 10^{-5}$ for the marginalized posterior value of ∇Z being consistent with the least squares value. Visually, at $f = 1$, both lines appear to describe the data fairly well. While there is slightly more variance in the data points around the flat profile predicted by the geostatistical model, this is not unexpected given the different ways that these two models were optimized – in the WLS approach, parameters for the line of best fit are chosen to minimize the scatter of all data points around the radial metallicity profile, whereas in our geostatistical model, data points are naturally expected to have some scatter around the large-scale trend (for $f = 1$, the most likely value for this scatter is $\sigma = 0.05$ dex).

When the WLS method was used, the metallicity gradient recovered for the highest resolution map ($\nabla Z = -0.0126$ dex/kpc) was within the range of metallicity gradients recovered for the other rebinned datacubes ($-0.013 \leq \nabla Z \leq -0.010$ dex/kpc), with a slight trend of the metallicity gradient becoming flatter at coarser resolutions, as is expected from a coarser spatial sampling. However, for the $\text{N}_2\text{S}_2\text{H}\alpha$ and Scal diagnostics, an increase in the WLS-reported metallicity gradient was seen at the finest resolutions. We discuss a possible interpretation for this effect in Section 4.2.1.

In Fig. 10, we plot the normalized prediction error (MAD/σ , as calculated using ten-fold cross-validation) of the geostatistical model, and the linear model with no small-scale component, against the resolution of the datacubes. We find that when a linear gradient model is used, a similar prediction error value of $\text{MAD}/\sigma = 0.75\text{--}0.85$ is seen at all resolutions, with a slight trend of the predictive accuracy decreasing as the resolution is made coarser. At all resolutions investigated, predictions from a geostatistical model were found to be more accurate than those from a metallicity-gradient model (although this improvement is only marginal for the coarsest resolution datacube with $f = 24$). For resolutions finer than ~ 330 pc (the scale at which adjacent pixels are separated by a distance less than the metallicity correlation scale, ϕ), the prediction accuracy of the geostatistical method improves rapidly as the resolution is improved. This highlights the power of this geostatistical modelling framework for being able to produce accurate predictive maps of

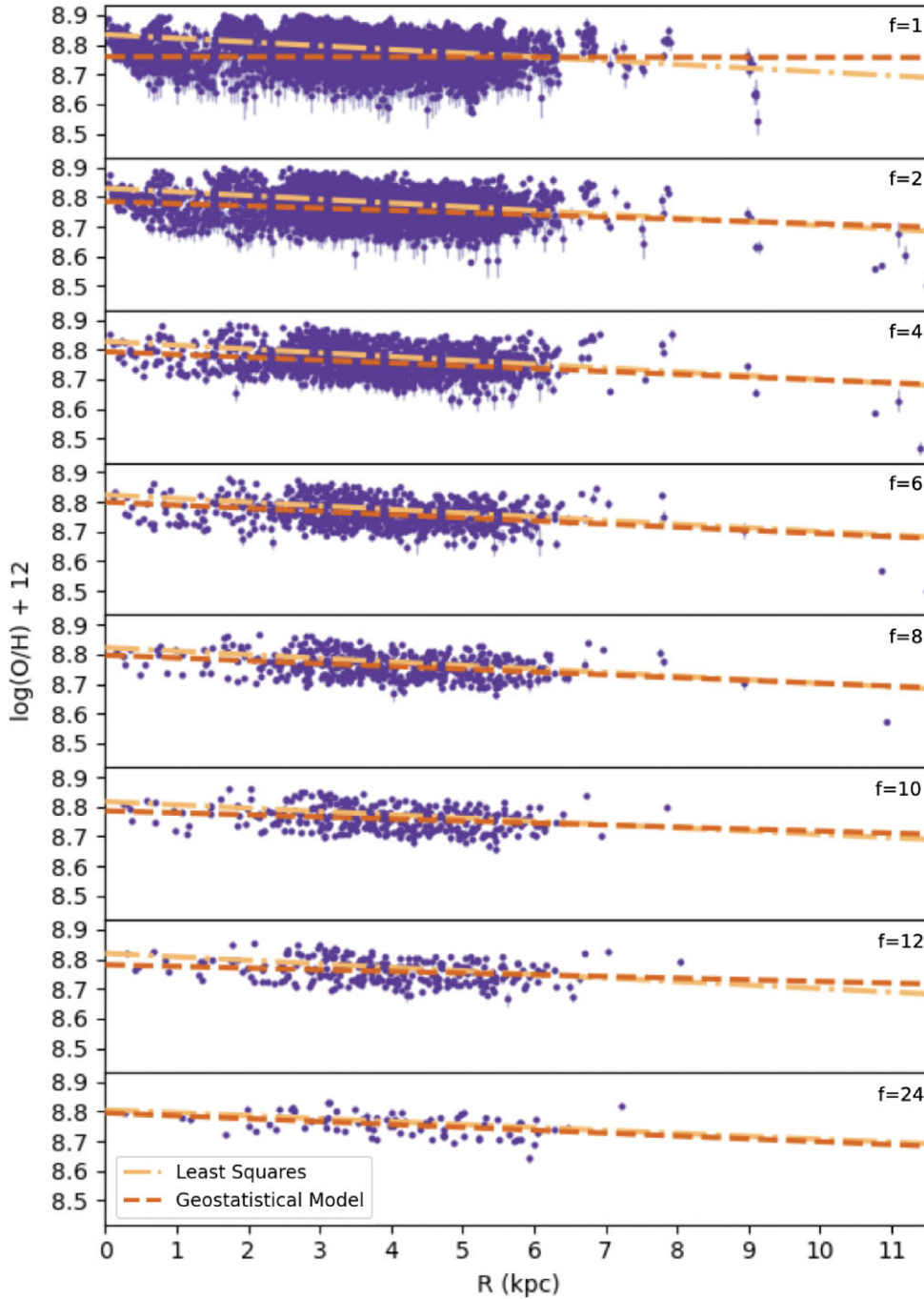


Figure 9. The radial metallicity profile of NGC 5236, computed using (i) our geostatistical methodology (red dashed lines) and (ii) the weighted least squares (WLS) algorithm (orange dashed lines) at different rebinned resolutions. When the WLS method is used, consistent values of the metallicity gradient are recovered for all binning factors f , with a slight tendency for the metallicity gradient to become flatter at the coarsest resolutions. For most of the resolutions tested, the radial metallicity profile recovered using our geostatistical approach agrees with the profile recovered using WLS. However, for the unbinned data ($f = 1$), the geostatistical method prefers a flatter metallicity gradient.

the 2D metallicity distributions of galaxies, especially for high-resolution datacubes.

We note that while the geostatistical model fit to the highest resolution data has a value of ∇Z that does not agree with the other models to 1σ , this model is still able to predict the metallicity of H II regions with more accuracy than any other model. This indicates that the predictive accuracy of our geostatistical model is not severely impacted by the different values of ∇Z recovered – the model with

a lower central metallicity and a flatter metallicity gradient is just as suitable for describing the observed data as a model with a stronger negative metallicity gradient.

4.2.1 Impact of spatial resolution on H II spaxel selection

There are many possible reasons why the metallicity gradients recovered with the finest resolution spaxels exhibit small, but statistically

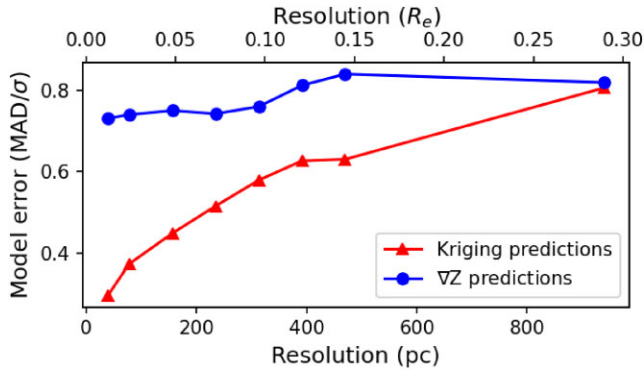


Figure 10. Normalized prediction error (median absolute deviation per standard deviation) for metallicity computed using the cross-validation method for NGC 5236 using a simple linear gradient (blue line) and a geostatistical model (red line). For all resolutions tested, predictions were more accurate when a geostatistical model including small-scale metallicity variations was included, with the greatest improvement seen at fine resolutions.

significant deviations compared to metallicity gradients measured at simulated coarser spatial resolutions.

First, it could be that the intrinsic radial trend in metallicity for these galaxies is not linear. High resolution metallicity maps have revealed breaks in a single metallicity profile that may not be visible in coarsely resolved data, potentially due to the influence of galactic bars (e.g. Chen et al. 2023). Flattening in metallicity profiles have been reported in the extended discs of spiral galaxies beyond R_{25} (e.g. Goddard et al. 2011; Rosales-Ortega et al. 2011), including for this galaxy (Bresolin et al. 2009). However, the data used in this study is limited to a maximum galactocentric distance of $\sim R_{25}$, where flattening has not been found in previous studies of this galaxy’s metallicity profile. Other studies have reported flatter metallicity profiles in the inner regions of galaxies (e.g. Belley & Roy 1992; Zinchenko et al. 2016), but we do not see this in NGC 5236 at any resolution (see Fig. 9).

The differences in inferred gradients could also be related to the selection of spaxels at each resolution. While we use a consistent H II spaxel selection method for all of our rebinned datacubes, the set of spaxels that are identified as being H II-dominated is not the same between the emission line maps with different resolution. This effect is visible in Fig. 9 – at $f = 2$, several spaxels are classified as H II-dominated at radii of ~ 11 kpc that have no analogues in the $f = 1$ map.

Additionally, (1) the H II region selection includes an empirically calibrated threshold from Kauffmann et al. (2003) based on Sloan Digital Sky Survey data that may not apply to highly resolved datacubes with resolutions finer than 100pc, and (2) the BPT diagnostic lines of Kewley et al. (2001) are computed to be the upper limits of the locations on the BPT diagrams that can be inhabited by H II regions – however, no guarantees are made that spaxels that are not ionized by H II emission must lie outside of this region. Therefore, if a statistically appreciable number of spaxels outside H II regions are instead flagged by the DIG diagnostics as H II dominated for finely-resolved data, this could lead to a bias in the metallicity gradient computed for high-resolution data.

To explore this hypothesis, we re-ran the analysis of Section 4.2, ensuring that the same selection of spaxels were used at each resolution. To do this, we followed a procedure similar to the one outlined in Poetrodjojo et al. (2019). First, using the native resolution data, for each metallicity diagnostic we define a mask containing only

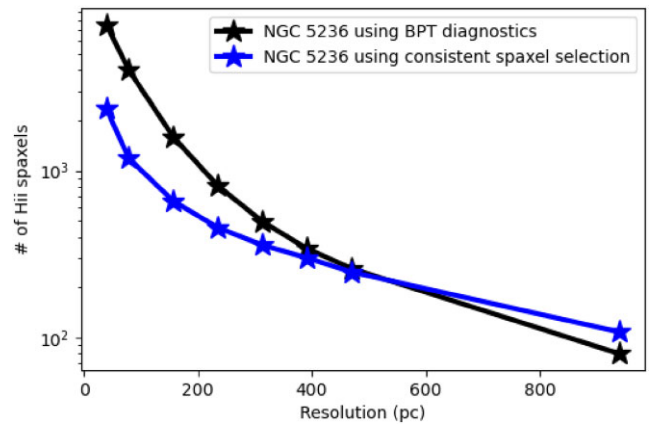


Figure 11. Number of data points available for NGC 5236 as a function of resolution, when H II spaxels are selected (i) using the BPT diagnostics of Kewley et al. (2001) and Kauffmann et al. (2003) together with the $H\alpha$ surface brightness cut of Zhang et al. (2017); and (ii) by selecting a sample of H II spaxels at native resolution using the line ratio $[S II]/H\alpha$, and rebinning this set of spaxels only to coarser resolutions. At the finest resolutions, the $[S II]/H\alpha$ -based cut removes many more spaxels than the BPT diagram based cuts. At coarser resolutions, however, more spaxels are retained, as binned spaxels that contain even one spaxel at native resolution that passes the data control cuts at native resolution are kept.

the spaxels that passed all data quality control cuts. When binning the data to coarser resolutions, only the spaxels that passed all data quality cuts at the finest resolutions were aggregated. For this test, we did not recompute line ratios for each rebinned spaxel with LZIFU. Instead, we simply summed the emission line fluxes of each spaxel within the mask contained within each binned spaxel, summing the errors in quadrature.

Similar to Poetrodjojo et al. (2019), and following M22, a harsher quality control cut for defining H II spaxels was used than the BPT diagram based methods. Based on the methods of Kaplan et al. (2016), we compute the fraction of light in each spaxel that is coming from H II regions (C_{HII}) using the line ratio $[S II]/H\alpha$. In more detail, using the native resolution maps, we assumed that the 100 spaxels with the brightest $H\alpha$ flux were entirely ionized by H II regions ($C_{HII} = 1$), and the 100 spaxels with the faintest $H\alpha$ flux (that still have a S/N ratio greater than 3 for both $[S II]$ and $H\alpha$) were entirely ionized by DIG ($C_{HII} = 0$). We take the median values of $[S II]/H\alpha$ of each of these populations to be the typical values of $[S II]/H\alpha$ for H II regions and the DIG, respectively, and computed values of C_{HII} for all spaxels by linearly interpolating where the measured value of $[S II]/H\alpha$ falls between these two extremes. Spaxels with a value of $C_{HII} > 0.95$ are considered to be H II dominated and are included in our mask.³ Additionally, any spaxels with $S/N < 3$ in $H\alpha$, $H\beta$, or any emission lines used in the metallicity diagnostic were discarded.

Fig. 11 shows how the number of spaxels available for our flagship galaxy (NGC 5236) with our fiducial metallicity diagnostic (RS_{32}) changes as a function of resolution with this new H II classification scheme. We compare this with the number of spaxels retained using the BPT diagrams and surface brightness cut as well. We see that, for the native resolution map, the $[S II]/H\alpha$ -based cut is much stricter

³This threshold value of 0.95 is (i) intrinsically arbitrary, and (ii) stricter than the threshold value of 0.9 used in Poetrodjojo et al. (2019). Such a conservative choice is consistent with the analysis presented in M22, where a 0.9 threshold was applied to $\log([S II]/H\alpha)$, which for the TYPHOON emission line map of NGC 5236 is equivalent to a 0.95 threshold in $[S II]/H\alpha$.

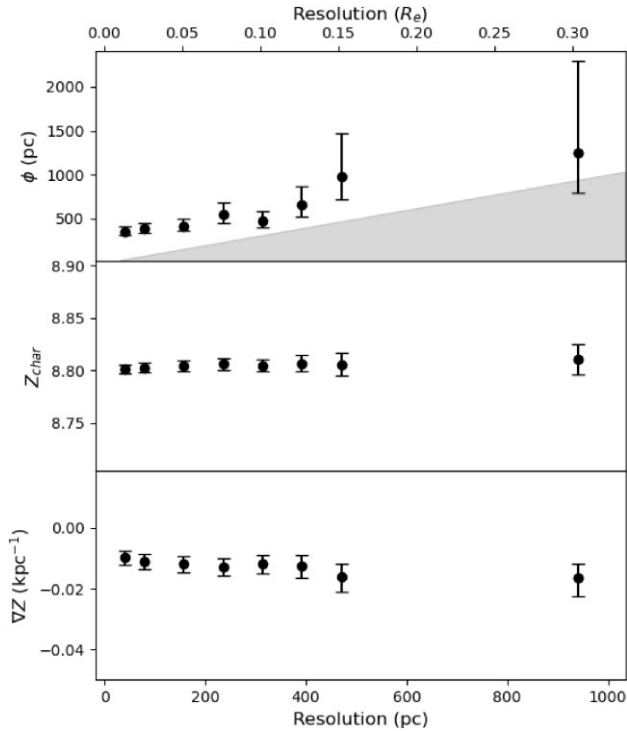


Figure 12. Parameters of our geostatistical model recovered with EMCEE after ensuring that the same set of spaxels are used at all resolutions. For this data, the value of ∇Z recovered does not become significantly flatter as the resolution is changed.

than the BPT diagram-based cut: while 7488 spaxels are classified as H II dominated at native resolution using the BPT diagram-based diagnostics, only 2351 pass the $[S\ II]/H\ \alpha$ -based cut and have $S/N > 3$ for $H\ \alpha$, $H\ \beta$, $[S\ II]\ \lambda\lambda 6717, 31$, and $[O\ III]\ \lambda 5007$. As the resolution is made coarser, the number of spaxels available for fitting decreases much less rapidly when the same set of spaxels are used at all resolutions. This may be due to an excess of spaxels being included at fine resolutions when the BPT-based spaxel selection criteria are used; or it could be due to spaxels being lost to DIG contamination at coarse resolutions.

Using a consistent selection of spaxels, but varying the resolution of the data, we then repeat our experiment of Section 4.2, exploring how our set of recovered parameters vary with resolution. We plot our results in Fig. 12. We find that in this case, the metallicity gradient varies much less than was found for the case where H II spaxels were selected using a BPT-based approach.

A similar result was seen in the investigations of Poetrodjojo et al. (2019). When DIG was excluded from the data at native resolution, metallicity gradients were found to be mostly constant with resolution. However, large variations in the metallicity gradient were found when different metallicity diagnostics were used, suggesting that systematic uncertainties in diagnostics are the most important limiting factor to achieve precise gradient measurements, rather than any effects resulting from having data with very fine spatial resolution.

4.3 The effects of aperture bias

In Sections 4.1 and 4.2, we assumed that all metallicities measured for H II spaxels are unbiased tracers of the underlying metallicities of the regions of the galaxy that they are drawn from. However,

there are reasons to consider this may not be the case. Spherically symmetric models of the interior structure of H II regions such as CLOUDY (Ferland et al. 2013) indicate that the ratio of emission lines within a H II region is expected to change as a function of the distance from the centre (see e.g. fig. 1 of Mannucci et al. 2021). For this reason, when the size of an IFS spaxel is smaller than the size of a H II region, the relative strengths of its emission lines are dependent not only on the metallicity within the spaxel, but also on the distance from the H II region’s centre to the spaxel’s centre, and on the relative sizes between the two. This may introduce an additional source of error in the metallicity estimated within each spaxel. By comparing the metallicities inferred from integrated DIG-free spectra to those that would be inferred from narrow slit spectroscopy, Mannucci et al. (2021) showed that in some circumstances aperture bias may have a stronger effect on recovered metallicities than DIG contamination. However, other studies with different methodologies have found the effects of aperture bias to be mild (e.g. Arellano-Córdova et al. 2022).

One step that can be added to an IFS data processing pipeline as a prophylactic measure against aperture bias is to first identify H II regions within a galaxy, and then integrate the emission from each spaxel within each region to get emission line ratios for each H II region. The line ratios for each region may then be converted into metallicities for each region. This ensures that no line ratio is artificially increased or decreased due to spatial sampling effects.

However, this method comes at some costs. First, by binning multiple adjacent spaxels into single regions, the spatial resolution of the data is decreased, and the number of available data points that can be used to fit a 2D metallicity model is lowered. Secondly, many H II region-finding algorithms make assumptions on the size, brightness, and geometry of H II regions that may not be valid (see e.g. Jin, Kewley & Sutherland 2022), yield different results at different spatial resolutions (Mast et al. 2014), and usually depend on parameters that must be hand-tuned by the user, making the results of a H II region search difficult to replicate (Lugo-Aranda et al. 2022).

In this context, a key question to ask is whether it is better to bin data into H II regions before beginning a geostatistical analysis, or to use higher resolution, possibly biased data from individual H II spaxels.

In Fig. 13, we show semivariograms for the small-scale metallicity variations within our fiducial galaxy (NGC 5236) using our fiducial metallicity diagnostic (RS_{32}), comparing results found when all spaxels are used (blue line) to results using data binned to 485 H II regions (red line), found using PYHIIEXTRACTOR that pass both BPT diagnostics (Section 3.1). We see that both semivariograms show the same shape, with the same flattening at ~ 0.4 kpc. Two other effects are also visible. First, the semivariogram produced using H II regions is more noisy. This is expected, because there are fewer data points available for the fitting – see Section 4.1 and Fig. 4. Secondly, the height of the semivariogram in the large-scale limit computed using H II regions is shorter than the semivariogram computed using H II spaxels. This, too, is expected, because the size of a H II region is larger than a H II spaxel, and so averaging over a larger region reduces the amount of variance in metallicity (see Fig. 7, and further discussion in Appendix C).

By comparing the height of the two semivariograms near zero separation, we see that the H II regions have far less uncorrelated variance than what is seen for H II spaxels. This observation has two possible interpretations. The first is that integrating over H II regions washes out small-scale metallicity variations that exist within an individual H II region. The second interpretation (which we consider more likely) is that metallicities estimated for integrated H II regions are more accurate than the metallicities estimated for individual

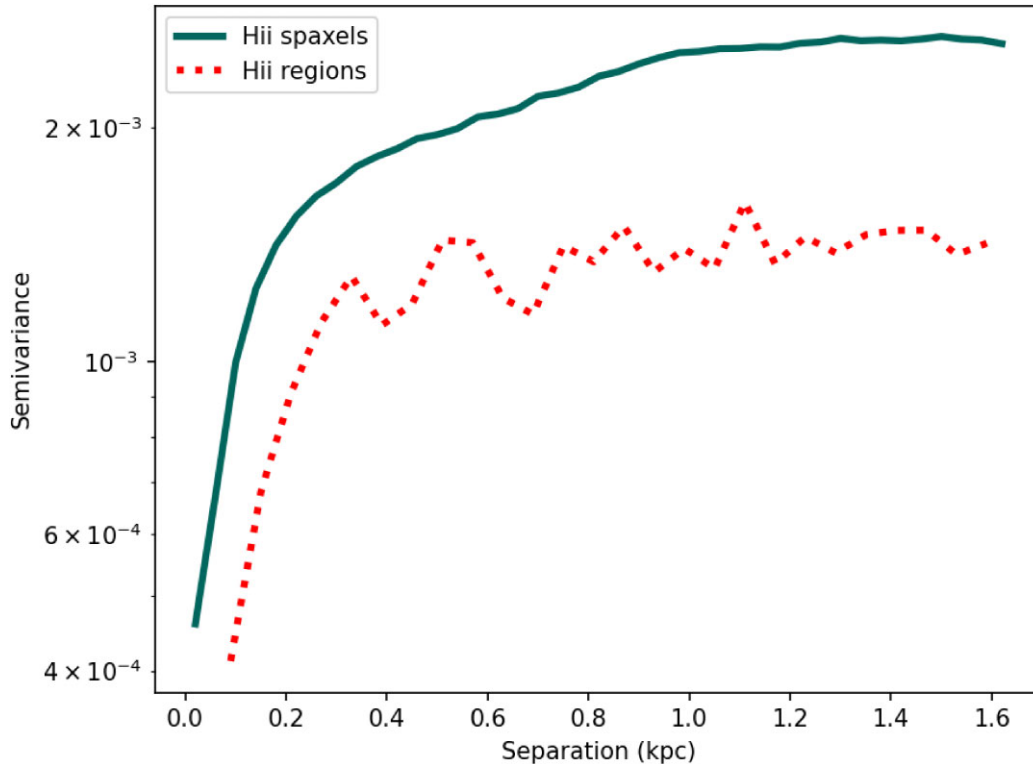


Figure 13. Comparing semivariograms calculated using H II dominated spaxels vs integrated H II regions found with PYHIIEXTRACTOR. We see that individual spaxels carry more noise than H II regions, and show a small but noticeable increase in variance from ~ 0.5 to 1 kpc that is not seen in the region-based semivariogram. The additional fluctuations in the region-based semivariogram are likely caused by a low number of data points compared to the spaxel-based semivariogram.

Table 2. Best-fitting parameter results for the metallicity model for NGC 5236 when results are computed using data from each spaxel associated with a H II region, compared to results when using integrated emission line ratios from H II regions found using PYHIIEXTRACTOR. In both cases, all spaxels/regions lie below the Kewley et al. (2001) and Kauffmann et al. (2003) demarcation lines in the S2- and N2-BPT diagrams. We find significant differences in the recovered values of two parameters for this galaxy with this diagnostic: the variance, which is understandable as H II regions are binned over a larger area than individual H II spaxels, naturally decreasing the variance in the metallicity between H II regions (see Appendix C), and the characteristic metallicity, Z_{char} , possibly reflecting a small systematic bias when this diagnostic is used on spaxels smaller than an individual H II region.

	H II spaxel fit	H II region fit
Z_{char}	8.76 ± 0.004	8.78 ± 0.004
∇Z (10^{-3} dex kpc $^{-1}$)	$-0.2^{+2.5}_{-2.4}$	-3.7 ± 2.7
ϕ (pc)	332^{+33}_{-28}	275^{+38}_{-32}
$\log_{10}(\sigma^2)$	-2.61 ± 0.04	-2.50 ± 0.04

H II spaxels. If this interpretation is correct, then by correcting for aperture bias, the error on the metallicity measurement of each spaxel is reduced. By comparing the height of these semivariograms near zero separation, the amount of variation reduction when this formulation is used could be estimated, potentially allowing the effects of aperture bias to be quantified for a range of commonly used metallicity diagnostics. We postpone such an analysis for a future study.

Overall, integrating spaxels into H II regions yields a metallicity map with fewer data points, coarser resolution, and more accurate

metallicity measurements. In Table 2, we show what effect this integration step has on the best-fitting values of Z_c , ∇Z , ϕ , and σ^2 recovered for this galaxy. We find that most of these parameters agree within their 68 per cent credible intervals. The variance is found to be significantly lower for H II regions, which is an understandable consequence of the coarser resolution of the binned H II region data compared to the unbinned H II spaxel data (see Appendix C). A 3.6σ difference is seen between the best-fitting values of Z_{char} for binned regions and individual spaxels, possibly pointing to a systematic error in metallicity measurements when this diagnostic is used on regions smaller than an individual H II region. However, the actual difference is very small, with the median values computed for regions and spaxels differing by only 0.02 dex. This implies that the effect of such a systematic bias, if it exists, is small enough to be ignored, particularly given the larger (0.7 dex) uncertainties associated with strong emission line ratio based metallicity diagnostics (Kewley & Ellison 2008).

For the other two parameters, there are some slight differences within the bounds of uncertainty. For the H II region data, $\nabla Z = -3.7 \pm 2.7 \times 10^{-3}$ dex kpc $^{-1}$. This is slightly more negative than the gradient found using native resolution H II spaxel data, but flatter than the gradients recovered for binned data, indicating that this data may be close to the limit of resolution at which the preferred models become those with no (or very weak) metallicity gradients, and variations within the galaxy is preferentially described through small-scale trends. Interestingly, the value of ϕ found for the H II region based data is lower than the value recovered when H II spaxels are used – but the size of this effect is not strong enough to be significant.

Based on this analysis, we conclude that binning H II spaxels into H II regions does reduce the uncertainty of the metallicity estimated at each data point, and does not significantly impact the results of a geostatistical model fit. As long as the data remains at a resolution finer than the expected size of small-scale variations within a galaxy, and the number of data points remains high ($\gtrsim 100$), we advise binning H II spaxels into H II regions before fitting a geostatistical model. However, if this is not possible, there is no cause for concern, as a model fit using unbinned H II spaxels will likely yield similar results.

A similar conclusion was reached by Vogt et al. (2017). They analysed the 2D abundance distribution of HCG 91c captured by MUSE in two ways: (i) spaxel-by spaxel, and (ii) binning the data into H II regions using the open-source Python fitting code contained in BRUTUS.⁴ They found that the metallicity gradient extracted was consistent when either spaxels or H II regions were used. Additionally, Vogt et al. (2017) noticed coherent, spatially localized variations in metallicity on sub-kpc scales when both methods were used, representing an amplitude fluctuation of 0.2 dex when computed from a spaxel-based analysis, and 0.15 dex when the data was binned into H II regions. Such fluctuations agree with what we find in our analysis, with the spatially correlated variation remaining after spaxels were binned, albeit with a slightly lower variance.

This result also shows that the fluctuations that we reveal with a semivariogram analysis are not caused by the varying ionization conditions within H II regions. Rather, they capture intrinsic variations in the metallicity of the ISM between H II regions.

5 RESULTS FOR NGC 6744 AND NGC 7793

In Section 4.2, we find that, based on the galaxy NGC 5236 and the RS_{32} diagnostic, the limiting resolution at which our geostatistical pipeline can accurately recover the statistical properties of small-scale metallicity fluctuations is ~ 350 pc. This value is comparable to the best-fit value of ϕ , the scale of these fluctuations. Because of this, it is unclear as to whether this resolution limit depends on the size of the intrinsic fluctuations in NGC 5236, or is constant over a wider galaxy sample.

We aim to make a qualitative assessment of the robustness of this resolution limit. For this, we expand our analysis to two other galaxies – NGC 6744, an intermediate spiral galaxy at a distance of 11.6 Mpc, and NGC 7793, a flocculent spiral galaxy 3.6 Mpc away. Each of these galaxies was binned to increasingly coarser resolutions, using values of $f = 1, 2, 4, 6, 8, 10, 12$. This corresponds to physical resolutions ranging from 93 to 1113 pc ($0.012 - 0.14R_e$) for NGC 6744, and 28 to 346 pc ($0.016 - 0.19R_e$) for NGC 7793. For NGC 7793, a binning with $f = 24$ (corresponding to a physical resolution of 691 pc) was also attempted, but only 31 spaxels satisfied our data cuts at this resolution, preventing a meaningful geostatistical model from being fit (c.f. Section 4.1).

At each resolution, a geostatistical model was fit to each galaxy using EMCEE. We show the median, 16th and 84th percentiles from our posterior distributions of ϕ , Z_{char} , and ∇Z for NGC 6744 at each resolution in Fig. 14, and for NGC 7793 in Fig. 15, using RS_{32} as our metallicity diagnostic. We also show the predictive accuracy of our geostatistical models using ten-fold cross-validation for NGC 6744 in Fig. 16 and for NGC 7793 in Fig. 17. We compare the metallicities predicted for each spaxel via universal kriging to the measured metallicities of each spaxel, comparing it against the

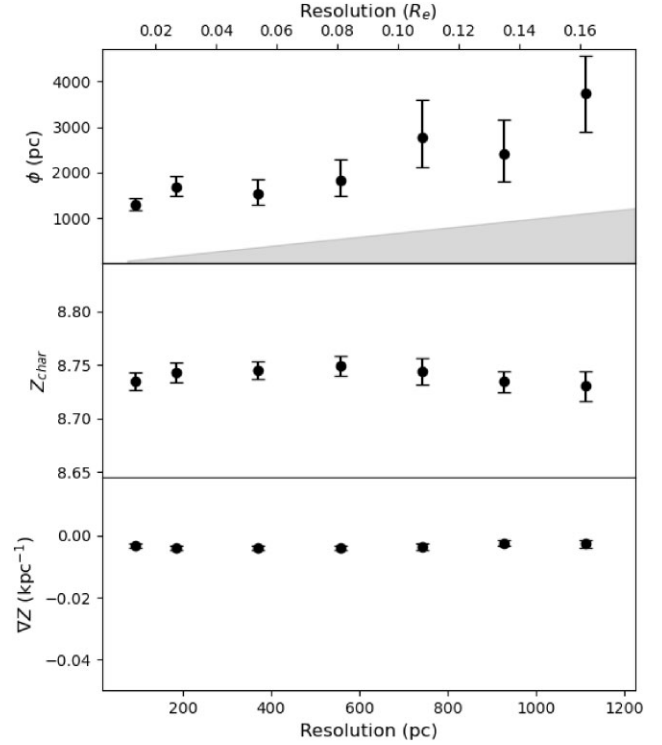


Figure 14. Best-fitting values, with uncertainty, of the metallicity correlation scale ϕ , the characteristic metallicity Z_{char} , and the metallicity gradient ∇Z for NGC 6744 as a function of the spatial resolution of metallicity maps. The grey-shaded region in the top panel represents the region where ϕ is less than or equal to the spaxel size.

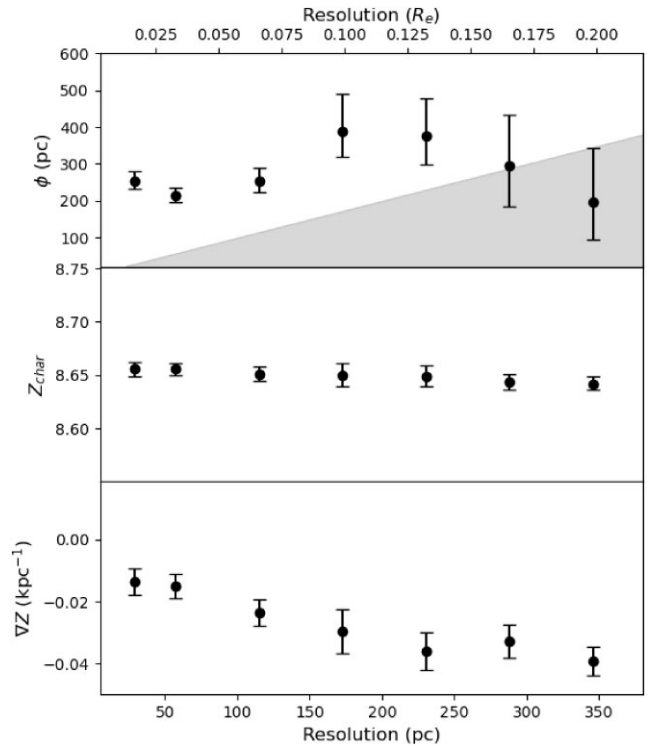


Figure 15. As in 14, but for NGC 7793.

⁴<https://fpavogt.github.io/brutus/>

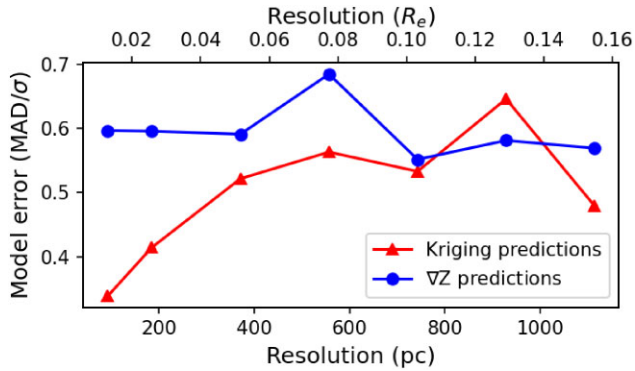


Figure 16. Normalized prediction error (median absolute deviation per standard deviation) of the predictions for metallicity computed using the cross-validation method for NGC 6744 using a simple linear gradient (blue line) and a geostatistical model (red line). When the small-scale structure can be resolved accurately, predictions from a geostatistical model are more accurate. However, at resolutions coarser than ~ 750 pc ($\sim 0.1R_e$), kriging predictions are no more accurate than predictions from a linear model with no small-scale component.

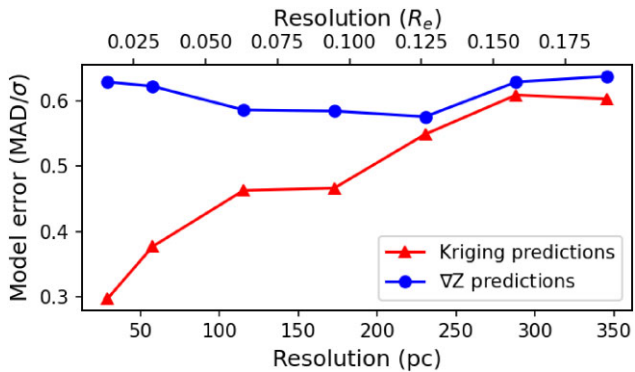


Figure 17. Normalized prediction error (median absolute deviation per standard deviation) of the predictions for metallicity computed using the cross-validation method for NGC 7793 using a simple linear gradient (blue line) and a geostatistical model (red line). When the small-scale structure can be resolved accurately, predictions from a geostatistical model are more accurate. However, at resolutions coarser than ~ 200 pc ($\sim 0.1R_e$), kriging predictions are no more accurate than predictions from a linear model with no small scale component.

predictive accuracy a simple linear gradient model fitted via the weighted least squares (WLS) method.

For NGC 6744, at native resolution, the most likely value of ϕ is $1.30^{+0.15}_{-0.12}$ kpc – much larger than the value of ϕ found for NGC 5236, even when NGC 5236 was binned to a resolution coarser than NGC 6744’s resolution. This indicates that correlations between metallicities exist on much larger scales in this galaxy than in NGC 5236. Here, we see that the correlation scale ϕ is fairly stable to increasing values of f until a resolution of 742 pc per pixel is reached, comparable to half the value of ϕ found at native resolution. After this point, the best-fitting values of ϕ become significantly higher, and the size of their error bars increases dramatically. This in turn affects the predictive accuracy of this model, which we plot in Fig. 16. After the point where the fit to ϕ stops being accurate, the geostatistical model becomes markedly worse, yet its predictive performance remains similar to a metallicity gradient model with no small-scale component.

For NGC 6744, the best-fitting values of the metallicity gradient and characteristic metallicity are stable for all resolutions tested, with all results consistent with each other within 1σ error bars.

For NGC 7793, at native resolution, the most likely value of ϕ is 254^{+24}_{-21} pc. Best-fitting values of ϕ are seen to deviate significantly from this value with large error bars when data with a resolution of $\gtrsim 170$ pc were used. This implies again that the result was stable to increases in resolution until a pixel becomes larger than $\sim \frac{1}{2}\phi$.

We show how the predictive accuracy of our model is affected by resolution for NGC 7793 in Fig. 17. At the two sharpest resolutions tested, the geostatistical model is able to predict metallicities at locations within the galaxy at a much greater accuracy than is possible with a simple linear gradient. At these resolutions, individual regions of enhanced star formation are well-resolved, and modelling small-scale deviations in metallicity captures much of the details of metallicity variability within this galaxy. At resolutions between 100 and 200 pc per spaxel, small-scale structure within this galaxy can still be marginally resolved and identified by our geostatistical pipeline, leading to models that, while not as successful at identifying local deviations from a large-scale metallicity trend, still exhibit a significant improvement in predictive power compared to a linear gradient model. When this galaxy is observed with resolutions coarser than ~ 200 pc per spaxel, the size of a pixel becomes comparable to the typical size of a small-scale metallicity fluctuation, and any significant benefit in predictive power from fitting a geostatistical model is lost.

Overall, this analysis shows that the resolution requirements for a geostatistical analysis to be valuable does indeed depend on the spatial extent (ϕ) of the small-scale variations within a galaxy. This makes coming up with general guidelines for the resolution requirements needed in order to study the internal metallicity distribution of galaxies using these geostatistical techniques difficult, especially when the typical size of small-scale fluctuations is not known in advance. However, we have noticed that the resolution at which geostatistical models begin to perform poorly often coincides with the resolution at which small-scale structure within a galaxy stops being apparent on a visual inspection.

This leads us to suggest the following ‘Golden Rule’ for when a geostatistical analysis is appropriate for IFS datacubes: when small-scale deviations from a linear metallicity gradient appear to exist *on visual inspection*, the size (in terms of both ϕ and σ^2) of these small-scale fluctuations can be captured accurately, and used to train predictive geostatistical models which can improve our understanding of small-scale metal mixing processes within galaxies. Our early results indicate that this is often achieved when the size of a resolution element is $\lesssim 0.1R_e$. However, we caution the reader that this guideline is only based on observations of three large local star-forming spiral galaxies, and it is not yet known if this relation will hold true for dwarf galaxies, or galaxies with other morphologies. At resolutions where small-scale fluctuations appear to exist but are not well-resolved (where ϕ is comparable to but smaller than the size of a pixel), geostatistical models offer some improvement in predictive power over a simple linear gradient model. Still, the best estimates of any parameters of the fit should be taken with some healthy scepticism, as poor resolution has been shown to bias ϕ to larger values, and lower σ^2 (see Appendix C). Finally, when the data is so coarse that no structure aside from a radial trend can be seen, there will not be enough data to fit a geostatistical model that is more informative than a metallicity gradient would be on its own.

6 DISCUSSION

From the results presented in Sections 4.2 and 5, we find that a resolution finer than $\sim \frac{1}{2}\phi$ is required to accurately recover the statistical properties of small-scale metallicity variations. From the literature (e.g. Kreckel et al. 2019, M21, M22), small-scale fluctuations in ISM metallicity are seldom seen on scales larger than 1kpc (however, see Williams et al. 2022)

For this reason, we do not consider it valuable to use this geostatistical methodology to analyse data products of ~ 1 kpc IFS surveys, such as SAMI (Croom et al. 2012), MaNGA (Bundy et al. 2015), or CALIFA (Sánchez et al. 2012). This is unfortunate, as these surveys contain a large number of galaxies spanning multiple orders of magnitude in mass and star formation rate, which would allow investigations into how the properties of small-scale metallicity variations are set by a galaxy’s global properties. However, there still exists a number of IFS surveys containing dozens of galaxies with spatial resolutions finer than ~ 250 pc. The combined catalogue AMUSING++ (López-Cobá et al. 2020) contains IFS observations of 635 galaxies observed with spaxel sizes of $0''.2$ and a typical seeing of $1''.0$. Approximately 40 per cent of galaxies in this sample are close enough to be resolved with a seeing finer than ~ 300 pc, making this a statistically significant legacy data set of IFS observations for which these geostatistical techniques could be applied. This data set also contains galaxies of different morphologies. Analysing such a diverse population of galaxies using a geostatistical pipeline could provide insights into how the different drivers of turbulence affect gas mixing in a galaxy throughout its lifetime.

Similarly, geostatistical analysis could be applied to very high resolution cosmological simulations of galaxy formation. Specifically, applications would be ideal for zoom simulations such as FIRE (Wetzel et al. 2023) and AURIGA (Grand et al. 2017), which simulate metallicity variations within galaxies on spatial scales as fine as ~ 2 pc. Such simulated data sets could readily be analysed with our geostatistical pipeline, in order to test the sub-grid physical prescriptions adopted in these simulations against observations, and thus both validate/falsify current choices and inform future theoretical models.

Extending a geostatistical analysis to the population of galaxies at cosmic noon and beyond will be possible with the next generation of extremely large thirty meter class ground-based telescopes. The European Extremely Large Telescope (E-ELT; Gilmozzi & Spyromilio 2007) and the Thirty Meter Telescope (TMT; Skidmore, TMT International Science Development Teams & Science Advisory Committee 2015) both include IFS instruments with resolutions of 4 milliarcseconds (HARMONI and IRIS, respectively) as part of their first generation suite of instruments. Under the Λ CDM cosmology of Planck Collaboration VI (2020b), this angular resolution corresponds to datacubes with a physical resolution finer than 35 parsec per pixel at *all redshifts*. Similarly, the IFS instrument GMTIFS planned for the Giant Magellan Telescope (GMT; Sharp et al. 2018) will have an angular resolution of 6 milliarcseconds, which corresponds to a physical resolution of 52 parsec per pixel or finer at all redshifts. Based on the resolution analysis presented here, any one of these instruments will produce data set suited for geostatistical analysis that will characterize the 2D chemical structure of galaxies at high redshift to a high level of detail. Comparing such galaxies to local ones, it should be possible to determine whether the process of metal mixing is the same in these chaotic starbursting systems as they are in the Universe today. In turn, this will contribute to understanding whether the processes that govern stellar feedback are constant throughout cosmic time. With existing high redshift metallicity diagnostics (e.g

Nakajima et al. 2022; Laseter et al. 2023; Sanders et al. 2023), based on the wavelength coverage of these instruments, such analysis will be possible for galaxies up to $z \approx 4$. Progressing beyond this redshift limit is possible if new rest-frame UV metallicity diagnostics for high redshift galaxies are developed.

Overall, our analysis highlighted the applicability and benefits of a geostatistical analysis. As with any novel diagnostic, the results obtained presents interesting aspects for further research. In particular, at the finest resolutions ($\lesssim 50$ pc per pixel), the best-fitting geostatistical models generally tended to fit weaker metallicity gradients than was seen for data at coarser resolutions. This behaviour is not intuitive, as other studies have shown that binning data to coarser resolutions leads to flatter recovered gradients for galaxies (Yuan, Kewley & Rich 2013; Mast et al. 2014; Acharyya et al. 2020). However, these studies have focused on simulating galaxy observations with resolutions far coarser than the range in which this effect was seen to be important.

One interpretation of this effect is that mean metallicity gradients of galaxies are an emergent property that comes about from averaging over small-scale fluctuations. This would imply that the global metallicity properties of galaxies are set by their small-scale gas physics; however, Baker et al. (2023) argue that kpc-scale features alone cannot be responsible for completely determining a galaxy’s metallicity. Another interpretation is that the metallicity gradient of a galaxy is not an intrinsic property of the galaxy itself, and that higher resolution data reveals significant departures from a linear trend that require a different model to produce a good fit.

We find that a flexible geostatistical model accounting for small-scale metallicity fluctuations within the data produces very good predictions for the metallicity of unknown regions within a galaxy, especially with high resolution data. Obtaining accurate metallicity predictions at specific points of a galaxy is important not only for studies of galaxy evolution and ISM physics, but also for cosmology. Accurate estimates of the metallicity at the site of many standard candles is required in order to calibrate their period–luminosity relations (e.g. Riepi et al. 2020; Bhardwaj et al. 2023). Therefore, the predictive accuracy of metallicity models is directly linked to the precision of local Universe measurements of the Hubble constant (Riess et al. 2022).

The circularly symmetric metallicity gradient model has been shown to be a good descriptor of the internal metallicity structure of local disc galaxies on the kpc-scale (e.g. Sánchez et al. 2014; Bresolin & Kennicutt 2015). For these galaxies, such a model is well-motivated by theories of inside-out galaxy evolution. However, its applicability to merging systems, dwarf galaxies, and galaxies that display large asymmetries in their surface brightness profile is less well motivated. At high-redshift, galaxies are known to be more clumpy and irregular than local galaxies (Conselice 2014), suggesting that models with more flexibility than a linear gradient model may be important for understanding the baryon cycle at cosmic noon and earlier epochs. Understanding the data quality and resolution required to make a meaningful fit to a galaxy’s metallicity map using a geostatistical model is an important first step to determining whether these models will be useful tools to analyse data from future IFS surveys of irregular systems and galaxies at high redshift.

In this work, the physical resolution of the data was defined to be the angular size of spaxels in pseudo-IFS data, times the distance to the galaxy. Distances to galaxies in the local universe can be highly uncertain. For example, we follow Leroy et al. (2019) in quoting the distance to NGC 6744 as 11.6 Mpc, but other references have found the distance to this galaxy to be as close as 7.6 ± 1.7 Mpc (Sorice

et al. 2014). Such large discrepancies in the distances to these galaxies make the values of ϕ much more uncertain than the values that we report – for example, adopting the distance from Sorce et al. (2014) would decrease the value of ϕ measured for NGC 6744 from 1.3 kpc to 852 pc. However, all results we have found on the relationship between the size of small-scale fluctuations and the resolution of the data would remain valid, as these two quantities scale with distance in a perfectly correlated way.

An alternative definition of the spatial resolution of an IFS datacube is based on the FWHM of the PSF for that data. For TYPHOON data, the PSF is several times smaller than the angular size of a spaxel. Because it is negligible at native resolution, we do not simulate the effects of having a large PSF that extends over many pixels at any coarser resolutions, for consistency. When the PSF is much larger than the size of a spaxel in an IFS observation, metallicity gradients appear flatter (Acharyya et al. 2020), and small-scale features may be blurred or lost. Techniques such as forward modelling can be used to model the effects of poor seeing on large-scale galaxy properties and correct for them (Metha et al. in preparation), and help distinguish between physical small-scale features and small-scale features caused by a known PSF profile. Additionally, correlations expected between nearby spaxels caused by instrumental effects could be directly incorporated into a geostatistical model by changing the form of $\epsilon(\vec{x})$ to capture this correlated structure.

Another important effect that was not directly controlled for in the series of resolution experiments presented in this work is the effect of altering the S/N of the data. Metallicity maps with coarser resolution will have a greater S/N than those at finer resolution due to binning. Presumably, this would affect the performance of a model-fitting procedure. We note that for all resolutions tested, the same cut of $S/N > 3$ on all lines used for metallicity diagnostics was enforced, ensuring that even at our finest resolution, no poor quality emission line fits are used to determine metallicities. We further note that geostatistical methods are suitable for distinguishing uncorrelated sources of noise, such as Poisson noise in a detector, from correlated signals of small-scale metallicity fluctuations. Thus, a geostatistical data analysis may be less impacted by having low S/N data than some other traditional statistical analysis techniques would be (see M21 for further details).

Finally, the inclination of a galaxy, i , is expected to make identification and classification of galaxy small-scale features more difficult, by effectively decreasing the resolution along one axis of a galaxy’s plane. Understanding how the geostatistical models presented in this work are affected by inclination is important to determine which systems these techniques are best applied to, and for choosing targets for future high-resolution IFS surveys. Because a galaxy can only be seen from one inclination angle, and small-scale properties of the ISM are not uniform between galaxies, mock-IFS observations of simulated galaxies are required to investigate this effect. We leave investigations on this subject for future work.

7 SUMMARY AND CONCLUSIONS

In this series of papers, we introduce a novel framework for interpretation and modelling of astronomical datacubes based on spatial statistics techniques. In M21, geostatistics was introduced as a useful tool for extracting additional data from high-resolution metallicity maps, allowing local deviations from large-scale metallicity profiles to be detected, classified, and compared to theoretical models of metal mixing in a galaxy’s ISM. In M22, we extended on this theme, showing how geostatistical models can be used to make accurate predictive maps of the metallicity structure of a galaxy

in 2-dimensions, including within regions where DIG contamination prevents the metallicity from being directly measurable. Here, we investigate how the capability of geostatistical methods to observe, model, and predict small-scale features in the metallicity maps of galaxies are affected by data quality, to increase the sample size of (real and simulated) galaxies over which these methods can be applied, and aid the design of future surveys that seek to provide detailed descriptions of the ISM of a wide variety of galaxies. In particular, we focus on how the best-fitting physical parameters and predictive power of a geostatistical model are affected by the resolution and completeness of an IFS-like observation. We summarize our main results below:

(i) We found that ~ 100 data points with known metallicity are required to train a geostatistical model of metallicity variations within a galaxy’s ISM. Below this value, noise due to small-number statistics makes computing a semivariogram very difficult, making it difficult to infer the presence of small-scale features. Furthermore, predictions from a geostatistical model perform no better than a linear gradient model. Increasing the number of data points available for a geostatistical analysis above this value increases the accuracy of kriging predictions, and decreases the uncertainty on the model’s parameters.

(ii) The resolution required to fit a geostatistical model that accurately captures small-scale features in the metallicity distributions of galaxies was found to depend on the size of those features. These features are typically of the order of 200–300 pc across, but vary from galaxy to galaxy. As a general rule, we find that the geostatistical data pipeline presented in this work is able to capture any small-scale features that are apparent on visual inspection; and from the galaxies explored in this work, we find that this tends to be achieved when the resolution is better than $\sim 0.1R_e$.

(iii) For coarsely resolved data, the best-fitting value of ϕ , the typical spatial extent of metallicity fluctuations, was found to depend on the resolution of the data. Therefore, best-fitting values of ϕ should be trusted if and only if the size of a pixel is less than $\phi/2$. In these cases, predictions from a geostatistical model vastly outperformed a metallicity gradient model.

(iv) Even when the resolution of the data was not sufficient to accurately constrain ϕ to its true value, kriging predictions were still found to be much more accurate than predictions from a linear gradient alone for data sets with $\gtrsim 100$ data points, and pixel sizes $< \phi$ where small-scale structure is marginally resolvable by eye. When fewer data points are available, or the resolution is coarser, predictions from kriging offered only a mild improvement over predictions from a linear metallicity gradient.

(v) A tradeoff exists between using integrated H II regions and individual H II-dominated spaxels for a spatially resolved metallicity analysis. Integrating spaxels into H II regions decreases the number of data points available and the resolution of the data, but also reduces the uncertainty of the metallicity of each data point. Ultimately, we found that both methods yielded consistent results. Therefore, we recommend binning data points into H II regions before a spatial analysis only when the average sizes of H II regions stays below $\sim \phi/2$, and the number of data points stays above ~ 100 .

Based on these results, there already exists a large legacy data set of local universe galaxy observations that have sufficient resolution and a sufficient number of data points for geostatistical models to be applied (López-Cobá et al. 2020). Key questions to answer include understanding how the metal mixing scale of galaxies varies with their size, morphology, star formation rate, and environment, to illuminate the interplay between small-scale gas physics and

galaxy evolution. Comparisons to zoom-in galaxy simulations will allow detailed models of ISM mixing to be directly compared to observational data. Future observations with high-resolution IFS instruments, such as the first generation IFS instruments planned for E-ELT, GMT, and TMT, will allow geostatistical methods to be applied to galaxies at redshifts up to $z \sim 4$, revealing how turbulent mixing processes within galaxies have changed over cosmic time.

ACKNOWLEDGEMENTS

We would like to thank the anonymous referee, whose comments helped to improve the quality of this paper. BM acknowledges support from an Australian Government Research Training Program (RTP) Scholarship. This research is supported in part by the Australian Research Council Centre of Excellence for All Sky Astrophysics in 3 Dimensions (ASTRO 3D), through project number CE170100013. BM would like to thank Prof. Tommaso Treu for his mentorship and advice on using MCMC methods and for hosting a wonderful extended research visit at UCLA. The authors extend their thanks to Alejandra Lugo-Aranda for the advice on using PYHIIEXTRACTOR, and to Dr Kathryn Grasha for the advice on using HIIPHOT and other helpful conversations. This research was conducted on Wurundjeri land.

DATA AVAILABILITY

The TYPHOON data will be publicly released in Seibert et al. (in prep). Prior to this release, the TYPHOON data are available upon reasonable request to Andrew Battisti. Any further data products created for this work are available from the corresponding author (BM) upon reasonable request. A publically available Python package containing implementations of the geostatistical techniques explored in this series is currently in preparation (Metha et al., in preparation).

REFERENCES

- Acharyya A., Krumholz M. R., Federrath C., Kewley L. J., Goldbaum N. J., Sharp R., 2020, *MNRAS*, 495, 3819
- Andrews S., Fryer C., Even W., Jones S., Pignatari M., 2020, *ApJ*, 890, 35
- Arellano-Córdova K. Z. et al., 2022, *ApJ*, 935, 74
- Baker W. M. et al., 2023, *MNRAS*, 519, 1149
- Baldwin J. A., Phillips M. M., Terlevich R., 1981, *PASP*, 93, 5
- Battisti A. J., Calzetti D., Chary R. R., 2017, *ApJ*, 851, 90
- Bec J., Khanin K., 2007, *Phys. Rep.*, 447, 1
- Belfiore F. et al., 2017, *MNRAS*, 469, 151
- Belley J., Roy J.-R., 1992, *ApJS*, 78, 61
- Berg D. A., Skillman E. D., Garnett D. R., Croxall K. V., Marble A. R., Smith J. D., Gordon K., Kennicutt R. C., Jr, 2013, *ApJ*, 775, 128
- Bhardwaj A. et al., 2023, *ApJ*, 944, L51
- Boissier S., Prantzos N., 1999, *MNRAS*, 307, 857
- Bresolin F., Kennicutt R. C., 2015, *MNRAS*, 454, 3664
- Bresolin F., Ryan-Weber E., Kennicutt R. C., Goddard Q., 2009, *ApJ*, 695, 580
- Bryant J. J. et al., 2015, *MNRAS*, 447, 2857
- Bundy K. et al., 2015, *ApJ*, 798, 7
- Cardelli J. A., Clayton G. C., Mathis J. S., 1989, *ApJ*, 345, 245
- Chen Q.-H., Grasha K., Battisti A. J., Kewley L. J., Madore B. F., Seibert M., Rich J. A., Beaton R. L., 2023, *MNRAS*, 519, 4801
- Christensen C. R., Davé R., Brooks A., Quinn T., Shen S., 2018, *ApJ*, 867, 142
- Conselice C. J., 2014, *ARA&A*, 52, 291
- Cosens M. et al., 2018, *ApJ*, 869, 11
- Cressie N., 1990, *Math. Geol.*, 22, 239
- Cressie N., 1993, *Statistics for Spatial Data*, Revised edn. Wiley, New York
- Croom S. M. et al., 2012, *MNRAS*, 421, 872
- Curti M., Mannucci F., Cresci G., Maiolino R., 2020, *MNRAS*, 491, 944
- de Vaucouleurs G., de Vaucouleurs A., Corwin H. G., Jr, Buta R. J., Paturel G., Fouque P., 1991, *Third Reference Catalogue of Bright Galaxies*. Springer, New York
- Dopita M. A., Kewley L. J., Sutherland R. S., Nicholls D. C., 2016, *Ap&SS*, 361, 61
- Emsellem E. et al., 2022, *A&A*, 659, A191
- Erroz-Ferrer S. et al., 2019, *MNRAS*, 484, 5009
- Espinosa-Ponce C., Sánchez S. F., Morisset C., Barrera-Ballesteros J. K., Galbany L., García-Benito R., Lacerda E. A. D., Mast D., 2020, *MNRAS*, 494, 1622
- Falceta-Gonçalves D., Kowal G., Falgarone E., Chian A. C. L., 2014, *Nonlinear Process. Geophys.*, 21, 587
- Feigelson E. D., Babu G. J., 2012, *Modern Statistical Methods for Astronomy*. Cambridge Univ. Press, Cambridge
- Ferland G. J. et al., 2013, *RMxAA*, 49, 137
- Field A., Miles J., Field Z., 2012, *Discovering Statistics using R*. SAGE Publications, London, England
- Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, *PASP*, 125, 306
- Gelfand A. E., 2010, in Gelfand A. E., Diggle P. J., Fuentes M., Guttorp P., eds, *Handbook of Spatial Statistics*. Taylor and Francis Group, Boca Raton, FL, p. 517
- Gilmozzi R., Spyromilio J., 2007, *The Messenger*, 127, 11
- Goddard Q. E., Bresolin F., Kennicutt R. C., Ryan-Weber E. V., Rosales-Ortega F. F., 2011, *MNRAS*, 412, 1246
- Grand R. J. J. et al., 2017, *MNRAS*, 467, 179
- Grasha K. et al., 2022, *ApJ*, 929, 118
- Groves B. et al., 2023, *MNRAS*, 520, 4902
- Haffner L. M. et al., 2009, *Rev. Mod. Phys.*, 81, 969
- Heckman T. M., Armus L., Miley G. K., 1990, *ApJS*, 74, 833
- Ho I. T., Kudritzki R.-P., Kewley L. J., Zahid H. J., Dopita M. A., Bresolin F., Rupke D. S. N., 2015, *MNRAS*, 448, 2030
- Ho I. T. et al., 2016, *Ap&SS*, 361, 280
- Ho I. T. et al., 2017, *ApJ*, 846, 39
- Ho I. T. et al., 2018, *A&A*, 618, A64
- Ho I. T. et al., 2019, *ApJ*, 885, L31
- Hogg D. W., Villar S., 2021, *PASP*, 133, 093001
- Jarrett T. H., Chester T., Cutri R., Schneider S. E., Huchra J. P., 2003, *AJ*, 125, 525
- Jin Y., Kewley L. J., Sutherland R. S., 2022, *ApJ*, 934, L8
- Journel A., Huijbregts C. J., 1978, *Mining Geostatistics*. Academic Press
- Kaplan K. F. et al., 2016, *MNRAS*, 462, 1642
- Kauffmann G. et al., 2003, *MNRAS*, 346, 1055
- Kewley L. J., Ellison S. L., 2008, *ApJ*, 681, 1183
- Kewley L. J., Heisler C. A., Dopita M. A., Lumsden S., 2001, *ApJSS*, 132, 37
- Kewley L. J., Nicholls D. C., Sutherland R. S., 2019, *ARA&A*, 57, 511
- Kreckel K. et al., 2019, *ApJ*, 887, 80
- Kreckel K. et al., 2020, *MNRAS*, 499, 193
- Kumari N., Maiolino R., Belfiore F., Curti M., 2019, *MNRAS*, 485, 367
- Larson R. B., 1972, *Nature*, 236, 21
- Laseter I. H. et al., 2024, *A&A*, 681, A70
- Lee D., Gammie C. F., 2021, *ApJ*, 906, 39
- Leroy A. K. et al., 2019, *ApJS*, 244, 24
- Li Z., Krumholz M. R., Wisnioski E., Mendel J. T., Kewley L. J., Sánchez S. F., Galbany L., 2021, *MNRAS*, 504, 5496
- Li Z. et al., 2023, *MNRAS*, 518, 286
- López-Cobá C. et al., 2020, *AJ*, 159, 167
- Lugo-Aranda A. Z., Sánchez S. F., Espinosa-Ponce C., López-Cobá C., Galbany L., Barrera-Ballesteros J. K., Sánchez-Menguiano L., Anderson J. P., 2022, *RASTI*, 1, 3
- Mannucci F. et al., 2021, *MNRAS*, 508, 1582
- Mast D. et al., 2014, *A&A*, 561, A129
- Matheron G., 1955, *Annales des mines.*, 144, 50
- Matheron G., 1963, *Econ. Geol.*, 58, 1246
- Metha B., Trenti M., Chu T., 2021, *MNRAS*, 508, 489

- Metha B., Trenti M., Chu T., Battisti A., 2022, *MNRAS*, 514, 4465
 Moustakas J., Kennicutt R. C., Jr, 2006, *ApJ*, 651, 155
 Naab T., Ostriker J. P., 2017, *ARA&A*, 55, 59
 Nakajima K. et al., 2022, *ApJS*, 262, 3
 Persson S. E., Carr D. M., Jacobs J. H., 1990, *Exp. Astron.*, 1, 195
 Pillepich A. et al., 2018, *MNRAS*, 473, 4077
 Pilyugin L. S., Grebel E. K., 2016, *MNRAS*, 457, 3678
 Planck Collaboration V, 2020a, *A&A*, 641, A5
 Planck Collaboration VI, 2020b, *A&A*, 641, A6
 Poetrodjojo H., D’Agostino J. J., Groves B., Kewley L., Ho I. T., Rich J., Madore B. F., Seibert M., 2019, *MNRAS*, 487, 79
 Poetrodjojo H. et al., 2021, *MNRAS*, 502, 3357
 Reichardt Chu B. et al., 2022, *ApJ*, 941, 163
 Reynolds R. J., 1990, in Bowyer S., Leinert C., eds, Proc. IAU Symp. 139, The Galactic and Extragalactic Background Radiation. Kluwer Academic Publishers, Dordrecht, Boston, p. 157
 Riess A. G. et al., 2022, *ApJ*, 934, L7
 Ripepi V. et al., 2020, *A&A*, 642, A230
 Rosales-Ortega F. F., Díaz A. I., Kennicutt R. C., Sánchez S. F., 2011, *MNRAS*, 415, 2439
 Sánchez-Menguiano L. et al., 2016, *ApJ*, 830, L40
 Sánchez-Menguiano L. et al., 2018, *A&A*, 609, A119
 Sánchez S. F., 2020, *ARA&A*, 58, 99
 Sánchez S. F. et al., 2012, *A&A*, 538, A8
 Sánchez S. F. et al., 2014, *A&A*, 563, A49
 Sanders R. L., Shapley A. E., Topping M. W., Reddy N. A., Brammer G. B., 2024, *ApJ*, 962, 24
 Schönrich R., McMillan P. J., 2017, *MNRAS*, 467, 1154
 Searle L., 1971, *ApJ*, 168, 327
 Sharp R. et al., 2018, in Evans C. J., Simard L., Takami H., eds, Proc. SPIE Conf. Ser. Vol. 10702, Ground-based and Airborne Instrumentation for Astronomy VII. SPIE, Bellingham, p. 107021V
 Skidmore W., TMT International Science Development Teams, Science Advisory Committee T., 2015, *Res. Astron. Astrophys.*, 15, 1945
 Sorce J. G., Tully R. B., Courtois H. M., Jarrett T. H., Neill J. D., Shaya E. J., 2014, *MNRAS*, 444, 527
 Stein M., 1999, *Interpolation of Spatial Data: Some Theory For Kriging*. Springer-Verlag, New York
 Sutherland R., Dopita M., Binette L., Groves B., 2018, *Astrophysics Source Code Library*, record ascl:1807.005
 Tobler W. R., 1970, *Econ. Geogr.*, 46, 234
 Trapp C. W. et al., 2022, *MNRAS*, 509, 4149
 Vila-Costas M. B., Edmunds M. G., 1992, *MNRAS*, 259, 121
 Vogt F. P. A., Pérez E., Dopita M. A., Verdes-Montenegro L., Borthakur S., 2017, *A&A*, 601, A61
 Wang X. et al., 2022, *ApJ*, 938, L16
 Wetzel A. et al., 2023, *ApJS*, 265, 44
 Wikle C. K., Zammit-Mangion A., Cressie C. N. A., 2019, *Spatio-Temporal Statistics with R*. CRC Press, Taylor & Francis Group
 Williams T. G. et al., 2022, *MNRAS*, 509, 1303
 Yuan T. T., Kewley L. J., Rich J., 2013, *ApJ*, 767, 106
 Zaritsky D., Kennicutt R. C., Jr, Huchra J. P., 1994, *ApJ*, 420, 87
 Zhang K. et al., 2017, *MNRAS*, 466, 3217
 Zinchenko I. A., Pilyugin L. S., Grebel E. K., Sánchez S. F., Vílchez J. M., 2016, *MNRAS*, 462, 2715

APPENDIX A: ANALYSIS WITH ALTERNATIVE METALLICITY DIAGNOSTICS

Estimating metallicities from strong emission line diagnostics has always been fraught with systematics. Different metallicity diagnostics can give metallicity estimates that differ by as much as 0.7 dex (Kewley & Ellison 2008). Furthermore, metallicity gradients measured in galaxies using different metallicity diagnostics can be different as well (Poetrodjojo et al. 2021).

For this reason, care must be taken to ensure that the findings of this study are not dependent on the specific metallicity diagnostic used.

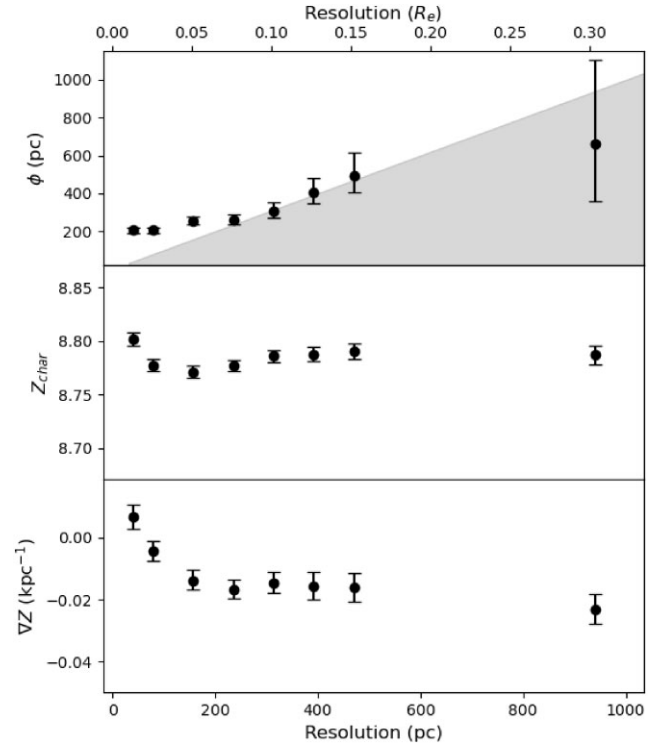


Figure A1. Medians, 16th, and 84th percentiles of the posterior distributions of the parameters of the geostatistical model recovered for NGC 5236, when $N_2S_2H\alpha$ is chosen to be the metallicity diagnostic. At the finest resolutions, a model with a metallicity gradient close to zero is preferred. A value of $\phi \approx 200$ pc is preferred when this diagnostic is used; and this value can be accurately recovered until the pixel size becomes comparable to this value.

To ensure our results are robust, we test three additional metallicity diagnostics, using three different emission line ratios, calibrated in a variety of ways. The first of these is the $N_2S_2H\alpha$ diagnostic, together with the theoretical calibration of Dopita et al. (2016) devised using the plasma modelling code MAPPINGS 5.0 (Sutherland et al. 2018). This diagnostic uses three emission lines: $[N\ II]\lambda 6583$, $H\alpha$, and the doublet $[S\ II]\lambda\lambda 6717, 6731$. This diagnostic has been shown to be largely insensitive to both dust extinction and variations in the ionization parameter.

Secondly, we consider the S-calibration (Scal hereafter) of Pilyugin & Grebel (2016). This empirical calibration was based on 313 reference H II regions with metallicities determined using the direct T_e -based method, and uses the emission lines $H\beta$, $[N\ II]\lambda 6583$, $[O\ III]\lambda 5007$, and the doublet $[S\ II]\lambda\lambda 6717, 6731$.

Finally, we examine how the results change when the O_3N_2 calibration of Curti et al. (2020) is used. This diagnostic was calibrated using exactly the same methods as our fiducial diagnostic, RS_{32} . However, unlike the other diagnostics discussed in this work, O_3N_2 has no dependence on sulphur, instead depending only on $H\alpha$, $H\beta$, $[O\ III]\lambda 5007$, and $[N\ II]\lambda 6853$. Thus, it provides an independent test against effects that could artificially raise or lower the relative strength of sulphur lines, such as DIG contamination (Haffner et al. 2009).

We show how the fitted values of ϕ , Z_{char} , and ∇Z vary with resolution for NGC 5236 when the $N_2S_2H\alpha$ diagnostic is used in Fig. A1. When this diagnostic is used, metallicity fluctuations are found to be correlated on a slightly smaller scale than was found when RS_{32} was used, with $\phi = 204 \pm 12$ pc at native resolution. A consistent value is found when $f = 2$, but for coarser resolutions

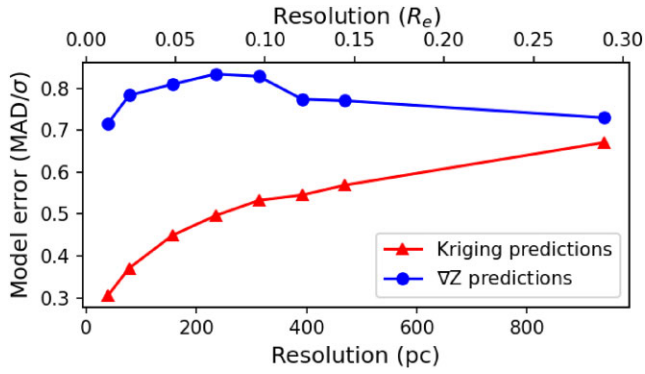


Figure A2. As in Fig. 10, but using $N_2S_2H\alpha$ as a metallicity diagnostic. We see that a geostatistical model fit still outperforms a linear gradient method at all resolutions tested.

with $f = 4$ and $f = 6$, this value increases slightly, to ~ 250 pc. At resolutions coarser than this, the recovered values of ϕ become increasingly high, as the size of a pixel becomes larger than the size of these small-scale fluctuations. We note a trend for the size of the fluctuations recovered to match the size of a pixel for four of the models fitted. This makes sense – our binning process removes all signal of correlation below the size of a single pixel, but some small-scale correlations are still observable in the data, so the model assumes that the scale of the correlation and the scale of a single pixel are the same.

It is unclear to us why the recovered size of metallicity fluctuations are different when this diagnostic is used, as compared to RS_{32} . None the less, from this sequence of model fits, we see that to accurately recover the value of ϕ , the size of a pixel must be smaller than $\sim \phi/2$ – the same result that is seen for the two other galaxies discussed in Section 5.

We also notice the same effect seen in Section 4.1, where the recovered value of ∇Z is close to zero at our two finest resolutions. At coarser resolutions than this, the value of ∇Z recovered is similar to, yet slightly steeper than, the metallicity gradient observed when our fiducial RS_{32} metallicity diagnostic is used. Disparities in the metallicity gradient recovered when different metallicity diagnostics are used have been seen before for this galaxy (e.g. Poetrodjojo et al. 2019), and the size of the difference in our study (~ 0.005 dex kpc^{-1}) is consistent with such expectations of variability across diagnostic tools.

In Fig. A2, we show the comparative predictive accuracy of our geostatistical model against a model that only contains a metallicity gradient at all resolutions tested. We see that our main results found when RS_{32} was chosen to be our metallicity diagnostic also hold for $N_2S_2H\alpha$: namely, (i) the geostatistical model outperforms a standard metallicity gradient model at all resolutions; (ii) the improvement in the fit is best seen for high-resolution data; and (iii) even when the recovered value of ∇Z is close to zero for finely resolved data, the predictions of the geostatistical model are still very accurate.

Our results when the $Scal$ diagnostic is used are very similar to the results found when $N_2S_2H\alpha$ is used. This is to be expected, as these two metallicity diagnostics have been found to be highly correlated (Groves et al. 2023). We show how the fitted values change as we change the resolution of our data in Fig. A3. At native resolution, $\phi = 222^{+18}_{-14}$ pc, consistent with the value found when $N_2S_2H\alpha$ is used. Consistent values of ϕ are observed until the size of a pixel becomes larger than 200 pc. After this point, ϕ is consistently estimated to be equal to the size of one pixel, until $f = 24$ where no small-

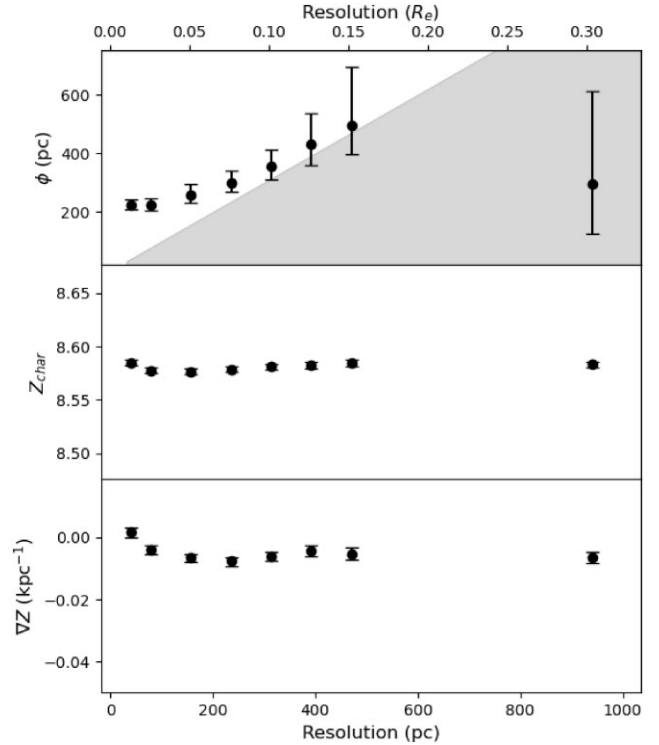


Figure A3. Medians, 16th, and 84th percentiles of the posterior distributions of the parameters of the geostatistical model recovered for NGC 5236, when $Scal$ is chosen to be the metallicity diagnostic. Values of ϕ recovered become unreliable when the pixel size approaches the size of the intrinsic fluctuations. At the finest resolution, the best-fitting value of ∇Z is close to zero, consistent with our other results.

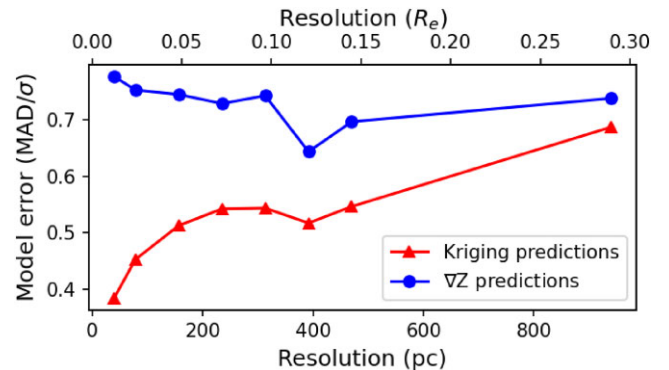


Figure A4. As in Fig. 10, but using $Scal$ as a metallicity diagnostic. Predictions from the geostatistical model always outperform those computed from the linear gradient model. The improvement is most significant at resolutions finer than ~ 200 pc, comparable to the value of ϕ measured for this galaxy using this diagnostic.

scale structure can be found, and the best-fitting value of ϕ is only recovered with very large uncertainties. As is found for the other two diagnostics tested, at the finest resolution, the geostatistical model prefers to fit a metallicity gradient that is close to zero.

We plot the predictive accuracy of a geostatistical model for NGC 5236 when this diagnostic is used compared to a metallicity-gradient only model in Fig. A4. We see the following results: (i) the geostatistical model always outperforms a linear fit; (ii) when the resolution is so coarse that ϕ cannot be accurately recovered,

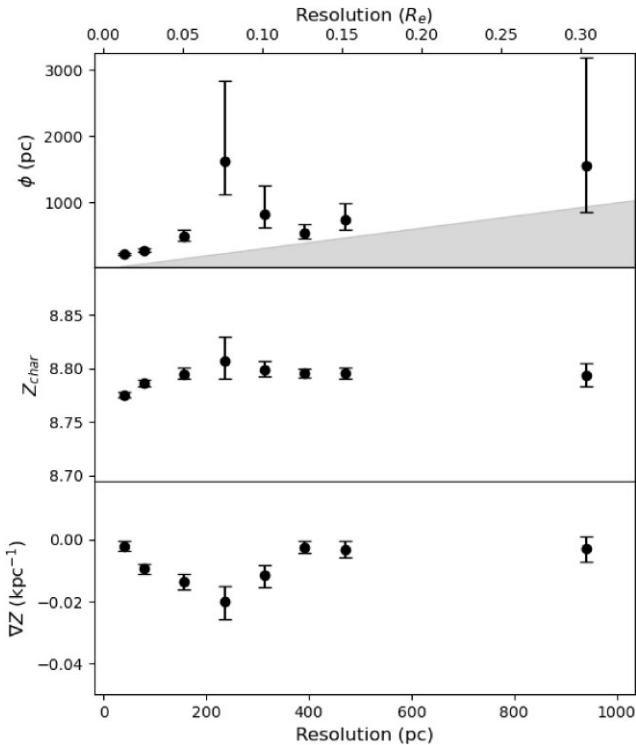


Figure A5. Medians, 16th, and 84th percentiles of the posterior distributions of the parameters of the geostatistical model recovered for NGC 5236, when O_3N_2 is chosen to be the metallicity diagnostic. When $f = 6$, small-scale structures stop becoming apparent in the data, as evidenced by the large value of ϕ . Around this point, the best-fitting value of ∇Z decreases and Z_{char} increases. For smaller and larger resolutions, the trend of ϕ increasing with resolution is similar to what is seen in other diagnostics.

the improvement over a linear gradient fit does not improve as the resolution of the data improves; (iii) when the data is of a fine enough resolution for ϕ to be recovered accurately, the predictive accuracy of the model continues to increase as the resolution is made finer, all the way down to native resolution; and (iv) having a recovered gradient close to zero at the finest resolution does not affect the predictive performance of this model.

Finally, we show the results when the O_3N_2 diagnostic is used. In Fig. A5, we plot the median, 16th, and 84th percentiles in the posteriors of each parameter computed from a Monte Carlo fit as a function of the resolution of the data input. We find that at native resolution, $\phi = 224^{+16}_{-15}$ pc, and ∇Z is flatter than is found at coarser resolutions, consistent with the other two diagnostics discussed in this Appendix. Results are similar when $f = 2$. However, an anomalous result is seen when $f = 6$, and possibly at the two resolutions closest to this resolution. For this particular bin factor, using this diagnostic, the best-fitting value of ϕ becomes very large, with large error bars: $\phi = 1.63^{+1.20}_{-0.51}$ kpc. This is larger than the value seen at any other resolution, including when $f = 24$. Non-monotonic trends are also seen in the recovered values of ∇Z and Z_{char} versus resolution when this diagnostic is used: at $f = 6$, the value of ∇Z reaches a minimum, and the (negatively correlated; $\rho = -0.57$) value of Z_{char} reaches a maximum. This indicates that for this particular resolution, with this particular diagnostic, the variance in the data can be preferentially described by a strong negative metallicity gradient, with no evidence for sub-kpc structures.

Further evidence for this interpretation is seen in Fig. A6. When $f = 6$, the median absolute difference between predicted metallicities

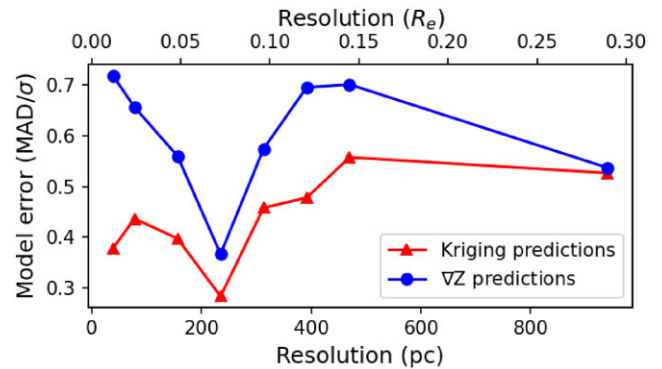


Figure A6. As in Fig. 10, but using O_3N_2 as a metallicity diagnostic. We find that when $f = 6$, a metallicity gradient alone provides an extremely good predictive model for the data. Even so, at all resolutions including when $f = 6$, the linear gradient model is outperformed by a geostatistical model.

from a linear gradient model and observed metallicities drops to 0.4σ – significantly smaller than the WLS model error seen when any other diagnostic is used to analyse NGC 5236 at any resolution, and half of the value of MAD/σ that is expected from Gaussian error. This indicates that for this combination of resolution and metallicity diagnostic, a linear gradient model does an excellent job of describing the metallicity variations within the data. Interestingly, the best-fitting metallicity gradient found using a WLS method does not change significantly between $f = 1$ and $f = 10$ ($-0.012 \leq \nabla Z \leq 0.09$ dex kpc^{-1}), indicating that this improvement in the performance of the model at this resolution is not caused by a more accurate recovery of the parameters, but by some difference in the properties of the data when this particular binning is used.

Even at $f = 6$, when the data is very well-described by a simple linear trend, we find that our geostatistical model still outperforms the linear fit, with a lower MAD/σ than is found even for the native resolution O_3N_2 metallicity map. We do not see any similar behaviour of the O_3N_2 diagnostic producing highly linear metallicity maps at particular resolutions for the other two galaxies investigated in this work. Because of this, we do not speculate as to what aspects of the O_3N_2 diagnostic may be responsible for this unusual behaviour.

Clearly, there is much work left to be done in understanding how strong emission line based metallicity diagnostics calibrated on stacked, integrated galaxy spectra may be applied to sub-kpc regions of the ISM. Extending on the conclusions of Kewley & Ellison (2008), we caution the reader that relying on individual strong emission line diagnostics may lead to erroneous conclusions about the metallicity structure of a galaxy. Following Kewley, Nicholls & Sutherland (2019), we recommend using a combination of multiple (three or more) strong emission line diagnostics using different line ratios to achieve a robust understanding of a galaxy’s internal metallicity distribution.

APPENDIX B: A PRIMER ON GEOSTATISTICAL METHODS

To capture details on the small-scale structure of the metallicity data around a large-scale metallicity trend, this work resorts to tools and techniques from *geostatistics*, the field of mathematics and statistics that concerns itself with analysing stochastic processes that occur over a continuous spatial domain. This field is vast, having evolved over several decades since the seminal works of Matheron (1954),

and we do not attempt to cover a comprehensive review here.⁵ Rather, we introduce two important concepts that have been discussed extensively in the previous papers in this series: the *semivariogram*, a tool for exploratory data analysis and separating correlated signals from uncorrelated noise, introduced for astronomical applications in M21; and *universal kriging*, an interpolation method that takes as input a set of data points and a geostatistical model, and provides as output predictions of the value of a random field at a set of specified locations (discussed extensively in M22). Furthermore, we provide a comparison between the semivariogram, the two-point correlation function and the power spectrum – two popular tools in the analysis of spatial structure in astronomical data sets (e.g. Planck Collaboration et al. 2020a; Li et al. 2021).

B1 The semivariogram and related methods

To capture details on the small-scale structure of the metallicity data around the metallicity trend, we use the *semivariogram*, a classic tool for exploratory data analysis from the geostatistical literature (Matheron 1963). The semivariogram is defined as

$$\gamma(h) = \frac{1}{2} \text{Var}(Z(\vec{x}) - Z(\vec{y})), \quad (\text{B1})$$

where the variance is computed over all pairs of points \vec{x} and \vec{y} for which $h - \delta_h/2 \leq |\vec{x} - \vec{y}| \leq h + \delta_h/2$. Here, δ_h is the size of bins used to compute the semivariogram, which should be chosen to be large enough so that enough pairs are present at each separation of interest that a variance can be reliably computed, but as small as possible to retain high spatial resolution (e.g. Journé & Huijbregts 1978; Cressie 1993).

Intuitively, the semivariogram shows how the variance between data points increases as the separation between them increases. Three informative parameters can be read from a semivariogram: the *range*, which is the separation between data points at which the variance ceases to increase, indicative of the sizes of the largest substructures within the data; the *sill*, which is the height of the semivariogram after it has flattened off, revealing the total variance in the data; and the *nugget*, which is the height of the semivariogram at zero separation, equivalent to the amount of uncorrelated noise in the data. By comparing the height of the sill to the nugget, we can quantify the amount of correlated versus uncorrelated variance in the data, allowing the amplitude, in dex, of metallicity fluctuations to be inferred while ignoring the effects of uncorrelated instrumental and calibration errors – see M21 for further discussion and an illustration.

The semivariogram is closely related to the two-point correlation function, $\xi(h)$, defined to be the Pearson correlation coefficient for the value of Z at two points separated by a distance of h . Under the assumptions that the random field under investigation is stationary and isotropic, the two diagnostic functions are related by the following equation:

$$\gamma(h) = \sigma^2 (1 - \xi(h)), \quad (\text{B2})$$

where σ^2 is the overall variance of Z throughout the field. Unlike the two-point correlation function, the semivariogram is not normalized; therefore, the semivariogram is sensitive to the amplitude of metallicity fluctuations, whereas the two-point correlation function is not.

⁵For those interested in geostatistics who want to read further, we recommend Wikle, Zammit-Mangion & Cressie (2019) as an introductory text, or Cressie (1993) for a comprehensive reference.

Another related metric often used in astronomy (and particularly cosmology) for determining the covariance structure of a random field is the power spectrum, which is calculated by taking the Fourier transform of a random field and then determining the square of the magnitude of each frequency component. By the Wiener–Khinchin theorem, for particularly well-behaved random fields, this function captures the same details that the semivariogram captures, but with units of spatial frequency rather than separation. However, the formal requirements on the data quality to use Fourier methods to reveal the covariance structure of the random field in this way are prohibitively strict: the random field must have no edges and no missing data, with periodic signals, and homoscedastic (constant variance) Gaussian error throughout (Feigelson & Babu 2012). For analysis of the cosmic microwave background where these methods find their natural home, these requirements are readily satisfied. However, metallicity maps of galaxies are always finite fields; are heteroscedastic (not homoscedastic) due to galaxies being brighter (and therefore having higher S/N) in their central regions; and have large portions of missing data due to DIG contamination. We therefore recommend the semivariogram as a more natural approach for determining the covariance structure of galaxy properties such as metallicity.

B2 Universal kriging

One reason why geostatistical models are so powerful is that they are predictive. After a geostatistical model has been fit to noisy data, it can be used to predict what the true values of the variable in question are if the noise was not present. It can also be used to predict the values of variables at points where no measurement exists, or where measurements are unreliable. This can be valuable in the field of spatially resolved metallicity analysis, where the metallicity found for many data points cannot be trusted due to low S/N or contamination from diffuse ionized gas (M22).

Kriging refers to a general family of inference techniques that provide unbiased, linear predictions for spatial data sets (Stein 1999) that are optimal in the sense that they minimize the mean square prediction error. Many classes of kriging exist, including ordinary kriging (in which the data has a constant mean), simple kriging (in which the mean values of the data are not constant but follow a known trend), robust kriging (designed to be insensitive to outliers) and even Bayesian kriging (which provides predictions for data points by marginalizing over a posterior of possible model parameters). More details on all of these methods can be found in Cressie (1993). For our application, we use *universal kriging*, in which the mean values of the data at each location are known to vary in space, but the best parameters to describe this trend are not known a priori, and must be learned from the data.

The mathematics that powers universal kriging is not complicated, but several definitions are required to properly setup the mathematical framework that is needed to completely define this process. We therefore refer the reader to M22 for the necessary equations, and instead provide some intuition on how this kriging method works below.

In universal kriging, the value at an unknown location is inferred from (i) the best-fitting value of the large-scale trend, and (ii) the (uncertain) values measured at all other data points. Based on the geostatistical model fit, the value of the unknown data point is expected to be strongly correlated with the values of data points nearby it, and weakly correlated with data points that are further away. In the model that we fit in this study, two data points separated by a distance of ϕ are expected to be positively correlated, with

a correlation coefficient of $\rho = 0.37$ (equation 4). This reflects the principle that things in nature that are close to each other tend to have similar properties (Tobler’s First Law of Geography; Tobler 1970).

Unlike other prediction methods, in addition to producing a predicted value at each location, kriging methods also provide uncertainties associated with each prediction. The uncertainty in the universal kriging prediction is the sum of three components: (i) σ^2 , the variance of the data, as fit by the geostatistical model (equation 4); (ii) a term that represents the reduction in variance from the fact that the data point is known to be correlated to other nearby data points; and (iii) an additional factor that captures how the uncertainty of the parameters describing the global model increase the uncertainty of our estimate of the data value.

In the limiting case where the location at which a prediction is desired is very close to a measured data point, the value of the prediction becomes the measured value of that data point, and its uncertainty on the kriging prediction becomes the uncertainty on the measured value of that data point. In the other limiting case, when the prediction is made far from any known data, the predicted value of the data becomes the prediction provided by the large-scale model, and its uncertainty is equal to the uncertainty of the model, plus a factor that represents the degree to which data points are seen to be scattered around this mean trend. Between these two limiting cases, predictions come from some combination of local, uncertain measurements, and large-scale trends, with the geostatistical model of best-fitting controlling to what degree each of these factors contribute to each prediction.

APPENDIX C: VARIANCE VERSUS PIXEL SIZE

In Section 4.2, the total variance between H II spaxels was shown to depend on the resolution of the data, with lower resolution maps showing less intrinsic variance. Qualitatively, this result is expected. Lower resolution maps contain larger pixels, resulting in reported metallicities being averaged over a larger number of H II regions in each pixel. This causes all metallicities reported to be closer to the mean metallicity of the galaxy, reducing the variance of metallicity within the galaxy.

Quantitatively, however, this result is not so simple. How can we explain the ratios of the variance between the native resolution map and the maps with binning factor f ? Furthermore, if we had a finer resolution IFU observation of this galaxy, how much would we expect the variance in metallicity to increase? To answer these questions, we must consider the correlation structure of the underlying metallicity random field, $Z(\vec{x})$, and the size of the pixels used to observe the random field (known in the geostatistical literature as the *support* of the data).

To understand how the visible features of a random field change as the support is changed, we follow Gelfand (2010). For each map with binning factor f , the true value of the metallicity at a point, $Z(\vec{x})$, is not observed. Instead, only the average metallicity over an area A_f can be seen:

$$Z_f(\vec{x}_0) = \frac{1}{|A_f|} \int_{A_{f0}} Z(\vec{s}) d\vec{s}. \quad (\text{C1})$$

Here, A_{f0} is an areal patch centred on \vec{x}_0 , with area $|A_f|$. For our purposes, we model each area A_{f0} as a square patch of sky centred on \vec{x}_0 with side length equal to $f \times 39$ pc, recalling that 39 pc is the native resolution of our observations of NGC 5236.

The covariance of $Z_f(\vec{x})$ can be calculated from the covariance of $Z(\vec{x})$. We assume that the covariance between the metallicity at two points depends only on their separation: that is, that there exists

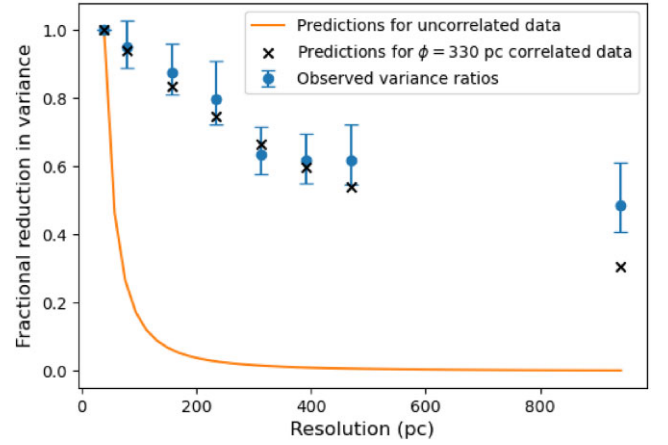


Figure C1. Blue points show the ratio of the height of the semivariograms pictured in Fig. 7 compared to the native resolution map with $f = 1$. Black crosses show the analytical expected reduction in variance when data is spatially correlated with a correlation scale length of $\phi = 330$ pc, following an exponential covariance structure. The orange line shows the naive reduction in variance expected when spatial correlations are not accounted for.

a function $C(\vec{x} - \vec{y})$ such that $\text{Cov}(Z(\vec{x}), Z(\vec{y})) = C(\vec{x} - \vec{y})$ for all points \vec{x}, \vec{y} in the galaxy. In this case, the covariance between two observations averaged over areas A_{f1} and A_{f2} centred on two point \vec{x}_1, \vec{x}_2 is given by

$$\text{Cov}(Z_f(\vec{x}_1), Z_f(\vec{x}_2)) = \frac{1}{|A_f|^2} \int_{A_{f1}} \int_{A_{f2}} C(\vec{s}_1 - \vec{s}_2) d\vec{s}_2 d\vec{s}_1. \quad (\text{C2})$$

If the metallicity of every spaxel was independent, we would expect the ratio of the variance in metallicity between the native resolution metallicity map and each map with degraded resolution to decrease as $1/N$, where $N = f^2$ is the number of native resolution spaxels contained within each degraded resolution spaxel. However, such a model predicts a much steeper decline in variance as the binning factor f increases than is seen in our data. Furthermore, as the resolution of the data is increased, under this model, we would expect the observed variance continue to increase indefinitely. This would be worrisome, as it would imply that the variance in metallicity seen within a galaxy at any finite resolution could never reflect the true intrinsic metallicity variability within a galaxy.

Instead, if we assume that the correlation between nearby values is approximately exponential, we get the following relation: $C(\vec{x} + \vec{y}) \approx \frac{1}{\sigma^2} C(\vec{x})C(\vec{y})$. Using this, we can relate the covariance of the averaged metallicities within each spaxel to the covariance of the intrinsic metallicity distribution by changing the variable \vec{s}_2 to $\vec{s}' = \vec{s}_2 - \vec{x}_2 + \vec{x}_1$:

$$\begin{aligned} \text{Cov}(Z_f(\vec{x}_1), Z_f(\vec{x}_2)) &= \frac{1}{|A_f|^2} \int_{A_{f1}} \int_{A_{f2}} C(\vec{s}_1 - \vec{s}_2) d\vec{s}_2 d\vec{s}_1 \\ &= \frac{1}{|A_f|^2} \int_{A_{f1}} \int_{A_{f1}} C(\vec{s}_1 - \vec{x}_2 + \vec{x}_1 - \vec{s}') d\vec{s}' d\vec{s}_1 \\ &= C(\vec{x}_1 - \vec{x}_2) \frac{1}{|A_f|^2 \sigma^2} \int_{A_{f1}} \int_{A_{f1}} C(\vec{s}_1 - \vec{s}') d\vec{s}' d\vec{s}_1. \end{aligned} \quad (\text{C3})$$

This shows that the covariance between the spatially averaged quantities $Z_f(\vec{x}_1)$ and $Z_f(\vec{x}_2)$ is, to first order, the same as the covariance between the intrinsic quantities $Z(\vec{x}_1)$ and $Z(\vec{x}_2)$, up to a factor of $K := \frac{1}{|A_f|^2 \sigma^2} \int_{A_{f1}} \int_{A_{f1}} C(\vec{s}_1 - \vec{s}') d\vec{s}' d\vec{s}_1$. This means that for a given model of the spatial correlation between data points, we can calculate this reduction in variance, K , as a function of pixel sizes.

We show the results of this calculation in Fig. C1 for NGC 5236 with our fiducial metallicity diagnostic, assuming an exponential covariance function with a correlation scale length of $\phi = 330$ pc (the best-fitting value calculated from the native resolution data; Section 4.2). We see that the reduction in variance as f is increased seen in the semivariograms displayed in Fig. 7 is consistent with the analytical predictions of this model. We also see that the reduction in variance is far less than would be assumed in a model in which no spatial correlations are accounted for.

Under this framework, we can also determine how much variance is lost when our native resolution map is used. If our resolution

was increased to 1 pc (for example, due to follow-up observations from ELT class instruments), our variance would only increase by a factor of 1.06, far smaller than the 39-fold increase that would be naively expected if spatial correlations were not accounted for. Because of this, we can be justified in saying that the variances we report for the small-scale metallicity structure in this galaxy have physical meaning, and are not merely functions of the resolution at which the galaxy is viewed.

This paper has been typeset from a $\text{\TeX}/\text{\LaTeX}$ file prepared by the author.