Mon. Not. R. Astron. Soc. **356,** 247–269 (2005) doi:10.1111/j.1365-2966.2004.08447.x

# The 2dF Galaxy Redshift Survey: stochastic relative biasing between galaxy populations

Vivienne Wild,[1⋆] John A. Peacock,[2] Ofer Lahav,[1,16] Edward Conway,[11] Steve Maddox,[11] Ivan K. Baldry,[9] Carlton M. Baugh,[6] Joss Bland-Hawthorn,[3] Terry Bridges,[3] Russell Cannon,[3] Shaun Cole,[6] Matthew Colless,[3] Chris Collins,[4] Warrick Couch,[5] Gavin Dalton,[7,15] Roberto De Propris,[3,17] Simon P. Driver,[14] George Efstathiou,[1] Richard S. Ellis,[8] Carlos S. Frenk,[6] Karl Glazebrook,[9] Carole Jackson,[14] Ian Lewis,[7] Stuart Lumsden,[10] Darren Madgwick,[12] Peder Norberg,[13] Bruce A. Peterson,[14] Will Sutherland[1] and Keith Taylor[8] (The 2dFGRS Team)

[1]*Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA*
[2]*Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ*
[3]*Anglo-Australian Observatory, PO Box 296, Epping, NSW 2111, Australia*
[4]*Astrophysics Research Institute, Liverpool John Moores University, 12 Quays House, Birkenhead L14 1LD*
[5]*Department of Astrophysics, University of New South Wales, Sydney, NSW 2052, Australia*
[6]*Department of Physics, University of Durham, South Road, Durham DH1 3LE*
[7]*Department of Physics, University of Oxford, Keble Road, Oxford OX1 3RH*
[8]*Department of Astronomy, California Institute of Technology, Pasadena, CA 91025, USA*
[9]*Department of Physics & Astronomy, Johns Hopkins University, Baltimore, MD 21118-2686, USA*
[10]*Department of Physics, University of Leeds, Woodhouse Lane, Leeds LS2 9JT*
[11]*School of Physics & Astronomy, University of Nottingham, Nottingham NG7 2RD*
[12]*Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA*
[13]*ETHZ Institut für Astronomie, HPF G3.1, ETH Hönggerberg, CH-8093 Zürich, Switzerland*
[14]*Research School of Astronomy & Astrophysics, The Australian National University, Weston Creek, ACT 2611, Australia*
[15]*Rutherford Appleton Laboratory, Chilton, Didcot OX11 0QX*
[16]*Department of Physics and Astronomy, University College London, London WC1E 6BT*
[17]*Astrophysics Group, Department of Physics, Bristol University, Tyndall Avenue, Bristol BS8 1TL*

Accepted 2004 September 23. Received 2004 September 16; in original form 2004 April 15

**ABSTRACT**

It is well known that the clustering of galaxies depends on galaxy type. Such *relative bias* complicates the inference of cosmological parameters from galaxy redshift surveys, and is a challenge to theories of galaxy formation and evolution. In this paper we perform a joint counts-in-cells analysis on galaxies in the 2dF Galaxy Redshift Survey, classified by both colour and spectral type, $\eta$, as early- or late-type galaxies. We fit three different models of relative bias to the joint probability distribution of the cell counts, assuming Poisson sampling of the galaxy density field. We investigate the non-linearity and stochasticity of the relative bias, with cubic cells of side $10 \leqslant L \leqslant 45$ Mpc ($h = 0.7$). Exact linear bias is ruled out with high significance on all scales. Power-law bias gives a better fit, but likelihood ratios prefer a bivariate lognormal distribution, with a non-zero 'stochasticity', i.e. scatter that may result from physical effects on galaxy formation other than those from the local density field. Using this model, we measure a correlation coefficient in log-density space ($r_{\rm LN}$) of 0.958 for cells of length $L = 10$ Mpc, increasing to 0.970 by $L = 45$ Mpc. This corresponds to a stochasticity $\sigma_{\rm b}/\hat{b}$ of 0.44 ± 0.02 and 0.27 ± 0.05, respectively. For smaller cells, the Poisson-sampled lognormal distribution presents an increasingly poor fit to the data, especially with regard to the fraction of completely empty cells. We compare these trends with the predictions of

⋆E-mail: vw@ast.cam.ac.uk

semi-analytic galaxy formation models: these match the data well in terms of the overall level of stochasticity, variation with scale and the fraction of empty cells.

**Key words:** methods: statistical – surveys – galaxies: distances and redshifts – galaxies: statistics – large-scale structure of Universe.



## 1 INTRODUCTION

The question of whether galaxies trace the matter distribution of the Universe has many implications for cosmology and galaxy formation theories. Since Hubble & Humason (1931) it has been known that galaxies of different type cluster differently, and as such it cannot be possible for all galaxies to trace the matter distribution exactly. This observation has been reconfirmed many times, traditionally by comparisons of the correlation functions of different subgroups. For example, early-type (or passive) galaxies are more strongly clustered than late-type (or actively star-forming) galaxies (e.g. Davis & Geller 1976; Dressler 1980; Lahav, Nemiroff & Piran 1990; Hermit et al. 1996; Norberg et al. 2002; Zehavi et al. 2002; Madgwick et al. 2003a) and luminous galaxies cluster more strongly than faint galaxies (e.g. Willmer, da Costa & Pellegrini 1998; Norberg et al. 2001, 2002; Zehavi et al. 2002, 2004).

Any difference in the distribution of galaxies relative to mass has become known as *galaxy bias*. This assumed a central importance in cosmology via the attempts to rescue the $\Omega_{\rm m} = 1$ universe after observations of cluster mass-to-light ratios suggested values closer to $\Omega_{\rm m} = 0.2$. Such a bias could occur if the galaxy formation efficiency were increased in overdense regions of space, the so-called 'high-peak scenario' (Davis et al. 1985; Bardeen et al. 1986). Although these efforts ultimately proved fruitless, understanding of bias remains important. In recent years much effort has been put into investigating galaxy bias through theory and numerical modelling, while observational results have been restricted by small survey volumes. With the advent of large galaxy redshift surveys such as the 2dF Galaxy Redshift Survey (2dFGRS: Colless et al. 2001, 2003), the Sloan Digital Sky Survey (SDSS: Strauss et al. 2002) and the Deep Extragalactic Evolutionary Probe (DEEP: Davis et al. 2003) it is becoming possible to quantify the galaxy distribution as never before, and provide detailed descriptions with which to compare theoretical and numerical models.

In principle, the form of bias should be derivable from the fundamental physical processes involved in galaxy formation; until we understand these, the bias remains a description of our ignorance. The simplest model of galaxy biasing is the linear biasing model: $\delta_{\rm g}(\boldsymbol{x}) = b\delta_m(\boldsymbol{x})$, where $\delta_{\rm g}$ is the galaxy overdensity perturbation, $\delta_m$ the mass overdensity perturbation and $b$ is a constant bias parameter. This model is unphysical for $b > 1$, as it allows negative densities. Alternative models in the literature fall into several basic classes: linear or non-linear, local or non-local, deterministic or stochastic. Locally biased galaxy formation (e.g. Coles 1993; Fry & Gaztanaga 1993; Scherrer & Weinberg 1998) depends only on the properties of the local environment, and the galaxy density is assumed to be a universal function of the matter density:

$$\delta_{\rm g} = f(\delta_m). \tag{1}$$

Because galaxies are discrete objects, this prescription is normally supplemented by the *Poisson clustering hypothesis*, in which galaxies are modelled as random events, the expectation number density of which is specified via $\delta_{\rm g}$. This model for discreteness can only be an approximation, but there is no simple alternative. We therefore assume Poisson sampling in what follows; for consistency, theoretical predictions are treated in the same way as the real data.

Non-local models (e.g. Bower et al. 1993; Matsubara 1999) arise when the efficiency of galaxy formation is modulated over scales larger than those over which the matter moves, for example by effects of quasar radiation on star formation. Stochastic bias (Pen 1998; Dekel & Lahav 1999, hereafter DL99) allows a range of values of $\delta_{\rm g}$ for a given $\delta_m$, above the Poissonian scatter caused by galaxy discreteness. Stochasticity is a natural part of non-local models, but some stochasticity is always expected to arise from physical processes of galaxy formation (Blanton et al. 1999). Throughout this paper we follow the general framework for non-linear stochastic biasing of DL99, in which the overdensity of one field can be related to that of a second field contained in the same volume of space through

$$\delta_1 = b(\delta_2)\delta_2 + \epsilon. \tag{2}$$

The scatter (stochasticity) in the relation is given by

$$\epsilon \equiv \delta_1 - \langle\delta_1\rangle. \tag{3}$$

In principle, the bias parameter $b$ can be any function of $\delta_2$; a constant value of $b$ and $\epsilon = 0$ represents deterministic linear biasing.

Galaxy bias is clearly of astrophysical interest in relation to an understanding of galaxy formation and evolution. Bias is also a major practical source of uncertainty in deriving cosmological constraints from galaxy surveys. Some particular examples are the measurement of $\beta = \Omega_{\rm m}^{0.6}/b$ (Peacock et al. 2001; Hawkins et al. 2003), where DL99 showed that stochastic effects could explain large discrepancies between results from different methods (for a review see Dekel & Ostriker 1999, table 7.2). Power spectrum measurements require constant bias as a fundamental assumption (Percival et al. 2001), and constraints placed on neutrino mass also assume scale-independent biasing (Elgarøy & Lahav 2003). Pen (1998) calculate the effect of non-linear stochastic bias on the measurement of the galaxy power spectrum on large scales, showing how the galaxy variance, bias and galaxy-dark matter cross-correlation coefficient can be calculated from velocity distortions in the power spectrum. The importance of biasing has increased still further with the release of the *WMAP* first year results (Spergel et al. 2003). In order to combine cosmic microwave background (CMB) and 2dFGRS data to give tighter constraints on cosmological parameters, a model for galaxy bias is required (Verde et al. 2003).

Three independent methods have been used to investigate galaxy biasing in the 2dFGRS catalogue. Lahav et al. (2002) combined pre-*WMAP* CMB and 2dFGRS data sets to measure the average bias over a range $0.02 < k < 0.15\,h\,{\rm Mpc}^{-1}$, concluding that galaxies are almost exactly unbiased on these scales. Verde et al. (2002) found the bias parameter to be consistent with unity over scales $0.1 < k < 0.5\,h\,{\rm Mpc}^{-1}$ through measurements of the 2dFGRS bispectrum.

A more direct method of studying the relation between mass and light is to map the dark matter using gravitational lensing. This field has made great progress in recent years, and it has been possible to measure not only the absolute degree of bias, but also its non-linearity and stochasticity (Fischer et al. 2000;

Hoekstra et al. 2002; Fan 2003; Pen et al. 2003). For example, Hoekstra et al. (2002) combine the Red-Sequence Cluster Survey and the VIRMOS-DESCART survey to find an average bias of $b = 0.71$ and a linear correlation coefficient of $r \simeq 0.57$ on scales of 1–2 $h^{-1}$ Mpc$^{-1}$. However, current weak-lensing measurements are dominated by non-linear and quasi-linear scales in the power spectrum, and it is not yet possible to say a great deal concerning bias in the very large-scale linear regime. This is of course the critical region for the interpretation of redshift surveys, where we want to know the relation between the power spectra of mass and light on $>100$ Mpc scales.

This question will be settled by future weak-lensing surveys. In the meantime, we can address a related simpler problem: the *relative* bias between subsets of galaxies. The morphological differences between galaxies and the link to their environments has been discussed for many decades as a potential clue to the nature and evolution of galaxy clustering (e.g. Spitzer & Baade 1951; Gunn & Gott 1972; Davis & Geller 1976; Yoshikawa et al. 2001). Modern galaxy redshift surveys allow us to split the galaxy population into a variety of subdivisions such as spectral type, colour and surface brightness. We can look at relative bias as a function of scale, and weighted by luminosity. This should yield important insights into the absolute degree of bias that may exist. Norberg et al. (2001) measured bias as a function of luminosity in the 2dFGRS, finding a bias relative to $L^*$ galaxies of $b/b^* = 0.85 + 0.15L/L^*$. We concentrate on the natural bimodality of the galaxy population, between red early types with little active star formation, and the blue late-type population (e.g. Baldry et al. 2003). Lahav & Saslaw (1992) measured bias as a function of morphological type and scale using the UGC, ESO and *IRAS* catalogues. The Las Campanas Redshift Survey has already provided some observational evidence against the linear deterministic model from splitting galaxies by their spectral types (Tegmark & Bromley 1999; Blanton 2000), and we present here a more extensive analysis of this type.

There are several complementary methods for the measurement of galaxy clustering, although most previous studies of the relative bias between galaxy types have concentrated on a relative bias parameter defined as the square root of the ratio of the correlation functions for the types under study. Madgwick et al. (2003a) used this method to measure the relative bias in the 2dFGRS, finding $b$ (passive/active) ranging from approximately 2.5 to 1.2 on scales of $0.2 < r < 20\,h^{-1}$ Mpc. However, even within such a large survey as the 2dFGRS the correlation functions become noisy beyond approximately $10\,h^{-1}$ Mpc. A second method is counts-in-cells, which can be directly related to the correlation function (Peebles 1980), and is optimized for the study of larger scales. It is this latter method that we employ in this paper. Conway et al. (2004) have also investigated the relative bias of different galaxy types using a counts-in-cells analysis of the 2dFGRS, but they use magnitude-limited samples, and consider only deterministic bias models, whereas our present analyses use volume-limited samples, and consider stochastic bias models. The counts-in-cells method has also been used to calculate the variance and higher-order moments of galaxy clustering in the 2dFGRS (Baugh et al. 2004; Conway et al. 2004; Croton et al. 2004a,b).

Many theoretical results on biasing from numerical models have also been reported. There are two main approaches to modelling galaxy distributions: semi-analytic (e.g. Kauffmann, Nusser & Steinmetz 1997; Somerville & Primack 1999; Benson et al. 2000) and hydrodynamic (e.g. Cen & Ostriker 1992; Blanton et al. 1999; Cen & Ostriker 2000; Yoshikawa et al. 2001). Comparisons are given by Helly et al. (2003) and Yoshida et al. (2002). Several studies have been made of galaxy biasing in these numerical simulations (e.g. Somerville et al. 2001; Yoshikawa et al. 2001), but none provide results in sufficient detail to allow an easy comparison with the 2dFGRS. We therefore analyse a large new semi-analytic calculation which is capable of yielding mock results that can be analysed in an identical manner to the real data.

In this paper we concentrate on a few aspects of relative bias mentioned above: splitting galaxies by spectral type and colour. We investigate the non-linearity, stochasticity and scale dependence of the biasing relation through comparison with three models. Section 2 summarizes the DL99 framework for biasing, and presents the bias models used in this paper. Section 3 describes the 2dFGRS catalogue, the derivations of the galaxies spectral types and colours and Section 4 explains the counts-in-cells method. In Section 5 we show the methods used for model fitting and error estimation. Section 6 gives the results and we compare our results with simulations in Section 7.

Throughout, we adopt a cosmological geometry with $\Omega_{\rm m} = 0.3$, $\Omega_{\rm v} = 0.7$ in order to convert redshifts and angles into three-dimensional comoving distances. We define our cells with $h = 0.7$, and all cell lengths are quoted in Mpc, instead of the standard $h^{-1}$ Mpc.

## 2 MODELLING RELATIVE BIAS

The simplest model for any bias (i.e. mass–galaxy, early–late, red–blue etc.) is that of *linear deterministic* bias: given a number of one type of object you can predict precisely (within Poisson errors) the number of the other type of object in the same region of space, and the relationship between the two numbers is linear. Recalling the familiar relation for the mass–galaxy distributions $\delta_{\rm g} = b\delta_m$, we can write $\delta_{\rm L} = b\delta_{\rm E}$ where $\delta_{\rm E}(\delta_{\rm L})$ denotes the overdensity of early-(late-) type galaxies in a volume of space. As described above, this empirical model can become unphysical in low-density regions. Considering the complex processes involved in galaxy formation, it would be surprising to find linear deterministic biasing to be true in all cases. Any reasonable physical theory in fact predicts non-trivial mass–galaxy biasing (Cole & Kaiser 1989) and simulations can also find biasing to be a complicated issue particularly on small scales (Cen & Ostriker 1992; Blanton et al. 1999; Somerville et al. 2001).

We investigate two potential improvements to the linear deterministic model. First, the bias could be non-linear, and some non-linearity is inevitable in order to 'fix' the unphysical properties of the linear model. Secondly, there may exist stochasticity (scatter beyond Poissonian discreteness noise), due to astrophysical processes involved in galaxy formation. DL99 presented a general framework to quantify these different aspects of biasing, and the following section summarizes their results.

### 2.1 A framework for non-linear, stochastic bias

We use the notation $f(\delta_{\rm E})$ and $f(\delta_{\rm L})$ to denote the one-point probability distribution functions (PDFs) for the fractional density fluctuations of early- and late-type galaxies. The fields $\delta_{\rm E}$ and $\delta_{\rm L}$ have zero mean by definition, and their variances are defined by

$$\sigma_i^2 \equiv \int_{-1}^{\infty} \delta_i^2 f(\delta_i)\,{\rm d}(\delta_i) \equiv \left\langle \delta_i^2 \right\rangle. \tag{4}$$

The joint underlying probability distribution of early- and late-type galaxies is given by

$$f(\delta_{\rm E}, \delta_{\rm L}) = f(\delta_{\rm E}) f(\delta_{\rm L}|\delta_{\rm E}) \tag{5}$$

$$= f(\delta_{\rm L}) f(\delta_{\rm E}|\delta_{\rm L}). \tag{6}$$

Both equations (5) and (6) should give the same outcome, but we have chosen to work with equation (5) to avoid unphysical linear biasing.

The natural generalization of linear biasing is given by

$$b(\delta_{\rm E})\delta_{\rm E} \equiv \langle \delta_{\rm L}|\delta_{\rm E}\rangle = \int f(\delta_{\rm L}|\delta_{\rm E})\delta_{\rm L}\,{\rm d}\delta_{\rm L}. \quad (7)$$

There are several useful statistics that can be used to investigate independently the fraction of non-linearity and stochasticity of a model or data. First, the mean biasing is defined by

$$\hat{b} \equiv \frac{\left\langle b(\delta_{\rm E})\delta_{\rm E}^2\right\rangle}{\sigma_{\rm E}^2}; \quad (8)$$

the non-linear equivalent of this is

$$\tilde{b}^2 \equiv \frac{\left\langle b^2(\delta_{\rm E})\delta_{\rm E}^2\right\rangle}{\sigma_{\rm E}^2}. \quad (9)$$

In each case the denominator is assigned such that linear biasing reduces to $b = \hat{b} = \tilde{b}$. The *random biasing field* is defined as

$$\epsilon \equiv \delta_{\rm L} - \langle \delta_{\rm L}|\delta_{\rm E}\rangle \quad (10)$$

and the statistical character of the biasing relation can be described via its variance, the *biasing scatter function*

$$\sigma_{\rm b}^2(\delta_{\rm E}) \equiv \frac{\left\langle \epsilon^2|\delta_{\rm E}\right\rangle}{\sigma_{\rm E}^2}. \quad (11)$$

The average biasing scatter is then given by

$$\sigma_{\rm b}^2 \equiv \frac{\langle \epsilon^2\rangle}{\sigma_{\rm E}^2}. \quad (12)$$

The purpose of this parametrization is to separate naturally the effects of non-linearity and stochasticity, allowing them to be quantified via the relations

$$\text{non-linearity} \equiv \tilde{b}/\hat{b} \quad (13)$$

$$\text{stochasticity} \equiv \sigma_{\rm b}/\hat{b}. \quad (14)$$

There are two further useful relations that are often quoted in the literature as measures of the bias parameter and stochasticity: the ratio of variances

$$b_{\rm var} \equiv \frac{\sigma_{\rm L}}{\sigma_{\rm E}} \quad (15)$$

and the linear correlation coefficient

$$r_{\rm lin} \equiv \frac{\langle \delta_{\rm E}\delta_{\rm L}\rangle}{\sigma_{\rm E}\sigma_{\rm L}} = \frac{\hat{b}}{b_{\rm var}}. \quad (16)$$

Both $b_{\rm var}$ and $r_{\rm lin}$ can be written in terms of the basic parameters given above, and both mix non-linear and stochastic effects. Non-parametric correlation coefficients can also be calculated directly from the data, although some method must be employed to account for shot noise. We refer the interested reader to DL99 for further details on the equations in this section.

Note that we work throughout with redshift-space overdensities. Redshift-space distortions are dependent on galaxy type due to the different clustering properties of early- and late-type galaxies. On non-linear scales the dominant effect is the finger-of-God stretching, but on the scales of interest to this paper we expect the linear $\beta$-effect to apply (Kaiser 1987). Averaging over all angles and including stochasticity between the galaxy and matter fields, we can write the redshift-space power spectrum, $P_{\rm s}$, as

$$\frac{P_{\rm s}}{P_{\rm r}} = \left(1 + \frac{2}{3}r\beta + \frac{1}{5}\beta^2\right), \quad (17)$$

where $\beta = \Omega_{\rm m}^{0.6}/b$, $b$ is the mass–galaxy bias, $r$ is the linear mass–galaxy correlation coefficient and $P_{\rm r}$ is the real-space power spectrum (Pen 1998; Dekel & Lahav 1999). The $\beta$-effect was measured for galaxies of different spectral class in the 2dFGRS by Madgwick et al. (2003a), obtaining $\beta_{\rm L} = 0.49 \pm 0.13$ and $\beta_{\rm E} = 0.48 \pm 0.14$. From these results and assuming $r = 1$ we obtain

$$\frac{P_{\rm s,E}}{P_{\rm s,L}} = 0.99\frac{P_{\rm r,E}}{P_{\rm r,L}}. \quad (18)$$

Although this suggests the effect is not large and currently insignificant within the errors, it is clear that in the case of zero stochasticity, redshift-space distortions will work to reduce the difference in the measured clustering between types. However, including a value of $r$ which is non-unity and dependent on galaxy type as suggested by simulations (e.g. Blanton et al. 1999), has a significant effect. For example, taking $r_{\rm L} = 0.8$ and $r_{\rm E} = 1.0$ increases the relative distortion from 0.99 to 1.04, where $r_{\rm L}$ ($r_{\rm E}$) is the linear correlation coefficient between the mass and late- (early-) type galaxy fields.

### 2.2 One-point probability distribution function

Given equation (5) we can split the model into two parts, first the distribution of early-type galaxies per cell, and secondly the biasing relation connecting the two distributions (see the following section). A standard description for the underlying probability distribution of a galaxy overdensity, $f(1+\delta)$, is lognormal (Coles & Jones 1991). Applying this for example to the early-type galaxies:

$$f(\delta_{\rm E})\,{\rm d}\delta_{\rm E} = \frac{1}{\omega_{\rm E}\sqrt{2\pi}}\exp\left(\frac{-x^2}{2\omega_{\rm E}^2}\right){\rm d}x, \quad (19)$$

where

$$x = \ln(1+\delta_{\rm E}) + \frac{\omega_{\rm E}^2}{2} \quad (20)$$

and $\omega_{\rm E}^2$ is the variance of the corresponding normal distribution $f[\ln(1+\delta)]$:

$$\omega_{\rm E}^2 = \left\langle [\ln(1+\delta_{\rm E})]^2\right\rangle. \quad (21)$$

The offset $\omega_{\rm E}^2/2$ is required to impose $\langle \delta_{\rm E}\rangle = 0$. If the lognormal distribution correctly describes the data, the variance of the overdensities, $\langle \delta_{\rm E}^2\rangle$, is related to the variance of the underlying Gaussian distribution by

$$\sigma_{\rm E}^2 \equiv \left\langle \delta_{\rm E}^2\right\rangle = \exp\left(\omega_{\rm E}^2\right) - 1. \quad (22)$$

In Section 6.5 we show that fitting a lognormal distribution directly to the data does not yield quite the same values for $\sigma_{\rm E}$ and $\sigma_{\rm L}$ as a direct variance estimate, but this does not affect our final results for stochasticity. On the largest scales, a lognormal distribution is completely consistent with the 2dFGRS data, and it provides a transparent and simple way to describe the density field.

### 2.3 Biasing models

#### 2.3.1 Deterministic bias: linear and power law

First, concentrating on deterministic bias, we can write the joint probability distribution function as

$$f(\delta_{\rm L}|\delta_{\rm E}) = \delta^{\rm D}[\delta_{\rm L} - b(\delta_{\rm E})\delta_{\rm E}], \quad (23)$$

where $\delta^{\rm D}$ is the Dirac delta function. This reduces directly to linear bias by setting

$$b(\delta_{\rm E})\delta_{\rm E} = b_{0,\mathcal{L}} + b_{1,\mathcal{L}}\delta_{\rm E}, \quad (24)$$

where the constraint $\langle\delta_{\rm L}\rangle = 0$ fixes $b_{0,\mathcal{L}} = 0$. A simple variation could be power-law bias

$$b(\delta_{\rm E})\delta_{\rm E} = b_{0,\mathcal{P}}(1+\delta_{\rm E})^{b_{1,\mathcal{P}}} - 1, \tag{25}$$

which avoids the negative density predictions of linear bias, and reduces to the linear biasing relation near $\delta = 0$. Rearranging equation (25), using the properties of lognormal distributions and the fact that $b_{1,\mathcal{P}} = \omega_{\rm L}/\omega_{\rm E}$, we find for the power-law bias that

$$b_{0,\mathcal{P}} = \exp\left[0.5\omega_{\rm E}^2\left(b_{1,\mathcal{P}} - b_{1,\mathcal{P}}^2\right)\right]. \tag{26}$$

For convenience we define

$$b_{\rm lin} = b_{1,\mathcal{L}} \tag{27}$$

and

$$b_{\rm pow} = b_{1,\mathcal{P}} \tag{28}$$

throughout the rest of this work.

### 2.3.2 *Stochastic bias: bivariate lognormal*

Returning to equation (7), we can introduce a broader function than the Dirac delta function of equation (23). An interesting class of model is when both $\delta$ fields form a bivariate Gaussian distribution, but this again becomes unphysical for $\delta < -1$ (DL99). It is, however, simple to cure this defect by assuming instead a *bivariate lognormal* distribution, for which the joint probability distribution is given by

$$f(g_{\rm E}, g_{\rm L}) = \frac{|V|^{-1/2}}{2\pi}\exp\left[-\frac{\left(\tilde{g_{\rm E}}^2 + \tilde{g_{\rm L}}^2 - 2r_{\rm LN}\,\tilde{g_{\rm E}}\,\tilde{g_{\rm L}}\right)}{2\left(1-r_{\rm LN}^2\right)}\right], \tag{29}$$

where $g_i = \ln(1+\delta_i) - \langle\ln(1+\delta_i)\rangle$ and $\tilde{g_i} = g_i/\omega_i$, with $i$ corresponding to early or late type. $\omega_i$ is related to the variance of the underlying Gaussian field $\ln(1+\delta_i)$ as for the one-point lognormal distribution:

$$\sigma_i^2 \equiv \left\langle\delta_i^2\right\rangle = \exp\left(\omega_i^2\right) - 1. \tag{30}$$

The correlation coefficient is

$$r_{\rm LN} = \frac{\langle g_{\rm E} g_{\rm L}\rangle}{\omega_{\rm E}\omega_{\rm L}} \equiv \frac{\omega_{\rm EL}^2}{\omega_{\rm E}\omega_{\rm L}} \tag{31}$$

and $V$ is the covariance matrix

$$V = \begin{pmatrix} \omega_{\rm E}^2 & \omega_{\rm EL}^2 \\ \omega_{\rm EL}^2 & \omega_{\rm L}^2 \end{pmatrix}. \tag{32}$$

Taking $f[\ln(1+\delta_{\rm E})]$ to be a Gaussian of width $\omega_{\rm E}$ and mean $-\omega_{\rm E}^2/2$ [i.e. $f(\delta_{\rm E})$ is distributed as a lognormal, equation (19)], the conditional probability distribution is

$$\begin{aligned} f(g_{\rm L}|g_{\rm E}) &= \frac{f(g_{\rm E}, g_{\rm L})}{f(g_{\rm E})} \\ &= \frac{\omega_{\rm E}}{(2\pi|V|)^{1/2}}\exp\left[-\frac{(\tilde{g_{\rm L}} - r_{\rm LN}\,\tilde{g_{\rm E}})^2}{2\left(1-r_{\rm LN}^2\right)}\right], \end{aligned} \tag{33}$$

i.e. the distribution of $\tilde{g_{\rm L}}|\tilde{g_{\rm E}}$ is a Gaussian with mean $r_{\rm LN}\,\tilde{g_{\rm E}}$ and variance $1 - r_{\rm LN}^2$.

As $r_{\rm LN} \to 1$, equation (33) reduces to a Dirac delta function, and this bivariate lognormal model reduces to the power-law bias model of equation (25). It is important to note that $r_{\rm LN}$ is not equal to the linear correlation coefficient $r_{\rm lin}$ of equation (16), which can differ from unity even if $r_{\rm LN} = 1$. In this sense, the lognormal parameters offer a cleaner separation of stochastic and non-linear effects. If stochasticity is present within the data, this model may provide an improvement over the deterministic biasing models. As observational data improve, it may become possible to constrain the relative biasing function to a greater extent; the current data are insufficient for such an analysis.

Analytic solutions exist for the mean biasing parameters and biasing scatter function given in Section 2.1 for this bivariate lognormal model. These relations are presented in the Appendix.

## 2.4 Including observational shot noise

It is not possible to measure the underlying probability distribution $f(\delta_{\rm E}, \delta_{\rm L})$ directly due to contamination of the observational data with noise, the dominant form of which is expected to be Poisson or 'shot' noise. This can be included in the models of the previous section by convolution with a Poisson distribution (Coles & Jones 1991; Blanton 2000). In this way the measured probability of finding $N_{\rm E}$ early-type galaxies and $N_{\rm L}$ late-type galaxies within a cell, $P(N_{\rm E}, N_{\rm L})$, can be compared with the models above. Accounting for shot noise in this way results in the models being less sensitive to outliers than equations (23) and (33).

Using equation (5) to combine the one-point PDF (19) with the conditional PDF (23 or 33), provides a model for the actual joint probability distribution function $f(\delta_{\rm E}, \delta_{\rm L})$. Convolution with a Poisson distribution then gives

$$\begin{aligned} P(N_{\rm E}, N_{\rm L}) = \int_{-1}^{\infty}\int_{-1}^{\infty} &\frac{\bar{N}_{\rm E}^{N_{\rm E}}(1+\delta_{\rm E})^{N_{\rm E}}}{N_{\rm E}!}{\rm e}^{-\bar{N}_{\rm E}(1+\delta_{\rm E})} f(\delta_{\rm E}) \\ &\times \frac{\bar{N}_{\rm L}^{N_{\rm L}}(1+\delta_{\rm L})^{N_{\rm L}}}{N_{\rm L}!}{\rm e}^{-\bar{N}_{\rm L}(1+\delta_{\rm L})} f(\delta_{\rm L}|\delta_{\rm E})\,{\rm d}\delta_{\rm E}\,{\rm d}\delta_{\rm L}, \end{aligned} \tag{34}$$

where $\bar{N}_{\rm E}$ ($\bar{N}_{\rm L}$) is the expected number of early- (late-) type galaxies in a given cell, allowing for completeness.

# 3 THE DATA: THE 2dFGRS

The 2dF Galaxy Redshift Survey, carried out between 1997 May and 2002 April, has obtained 221 414 good-quality galaxy spectra using the multi-object spectrograph 2dF on the Anglo-Australian Telescope. The main survey area comprises two rectangular strips of sky with boundaries ($09^{\rm h}\,50^{\rm m} < \alpha < 14^{\rm h}\,50^{\rm m}$, $-7.5^\circ < \delta < +2.5^\circ$) for the North Galactic Pole (NGP) and ($21^{\rm h}\,40^{\rm m} < \alpha < 03^{\rm h}\,40^{\rm m}$, $-37.5^\circ < \delta < -22.5^\circ$) for the South Galactic Pole (SGP), with a galaxy median redshift of $\bar{z} = 0.11$. At the median redshift, the physical size of the survey strips is $375\,h^{-1}$ Mpc long and the SGP and NGP regions have widths of 75 and $37.5\,h^{-1}$ Mpc, respectively (Peacock 2003). The input galaxies were selected from a revised and extended version of the APM galaxy catalogue (Maddox et al. 1990), and have a limiting extinction-corrected magnitude of $B_{\rm J} = 19.45$. Further details of the 2dFGRS can be found in Colless et al. (2001, 2003), and on the web at http://msowww.anu.edu.au/2dFGRS/.

For any structure analysis it is important to be aware of several problems that cause varying completeness over the survey region. Common to all similar surveys, some regions of the sky must be masked due to bright stars causing internal holes. Furthermore, due to the adaptive tiling algorithm employed to ensure an optimal observing strategy, the sampling fraction falls to as little as 50 per cent near the survey boundaries and internal holes due to lack of tile overlap. Subsequent reanalysis of the photometry of the APM galaxy catalogue has shown the survey depth to vary slightly with position on the sky. To account for this, we use a limiting corrected magnitude of $B_{\rm J} = 19.2$.

### 3.1 The volume-limited galaxy sample

It is well known that the luminosity of a galaxy is correlated with galaxy type. Therefore, in a flux-limited sample the fraction of early/late types varies with redshift, potentially complicating the analysis. Within any redshift survey the number density of objects drops substantially as we reach beyond the $L_*$ galaxy luminosity. The size of the 2dFGRS presents the option of studying volume-limited galaxy samples rather than the more usual flux-limited data sets of previous galaxy redshift surveys. By imposing a luminosity and redshift cut, volume-limited samples contain a representative sample of most galaxies over a large redshift range. Although some faint galaxies at low redshift are lost from the analysis, the sample selection effects are greatly simplified.

We use the publicly released data of 2003 June, containing a total of 221 414 unique galaxies with reliable redshifts, 192 979 of which have spectral classification. An absolute magnitude limit of $M_{B_J} - 5\log_{10}(h) \leqslant -19.0$ gives a representative sample of the local population, maximizing the number of cells, versus the number of galaxies in each cell. The absolute magnitude is given by $M_{B_J} = m - DM - K(z)$, where $m$ is the apparent extinction-corrected magnitude, $DM$ is the distance modulus and $K(z)$ is the $K$-correction. Setting a limiting survey magnitude of $m = 19.2$ allows for the varying survey depth with position on the sky (Colless et al. 2001), the $K$-corrections as a function of $\eta$ type are given in Madgwick et al. (2002). This gives a maximum redshift for our sample set by type 1 galaxies of $z_{\max} = 0.114$, and 48 066 galaxies in total. 46 912 of these have a spectral classification and 48 040 have measured colours. To account for the selection effects within the survey we use the publicly available redshift completeness masks. These are sufficient for galaxies classified by colour, but the spectral-type analysis introduces extra selection effects. A difference in completeness over a region of sky could occur, for example, when the spectra of a survey plate are of too poor quality to perform the spectral-type analysis, yet redshifts can be obtained. It is necessary to create separate masks to include these effects using the publicly available software.

### 3.2 Galaxy properties in the 2dFGRS

The 2dFGRS catalogue provides two methods of classification for comparison. First, the well-studied galaxy spectral type $\eta$, and secondly the photometric colours of the galaxies have recently been derived.

#### 3.2.1 Spectral type, $\eta$

The spectral type of the galaxies has been derived by a principal-component analysis (PCA), which identifies the most variable aspects of the galaxy spectra with no prior assumptions or template spectra (Folkes et al. 1999; Madgwick et al. 2002). The spectral type of the 2dFGRS galaxies is characterized by the value $\eta$, a linear combination of the first two principal components, derived in order to minimize the effect of distortions and imperfections in the 2dFGRS spectra. In effect, $\eta$ classifies galaxies according to the average emission and absorption-line strength in the spectrum. $\eta$ provides a continuous classification scheme, but for our purposes it is necessary to split the galaxies into two classes, at $\eta = -1.4$ as suggested in Madgwick (2003). Galaxies with $\eta < -1.4$ (type 1) are shown to be predominantly passive galaxies and those with $\eta > -1.4$ (types 2–4) predominantly star-forming (Madgwick et al. 2003b). The former are hence termed 'early type' and the latter 'late type'. The 2dFGRS catalogue contains 74 548 early-type and 118 424 late-type galaxies defined in this way.

One concern with using optical fibre spectra for this type of analysis are 'aperture effects', resulting from the fixed aperture of the fibres being smaller than the size of galaxies. This could result in, for example, only the bulge components of close spirals being observed. Such effects have been studied in detail by Madgwick et al. (2002), and no systematic bias found. One possible explanation for this is the poor seeing present at the Anglo-Australian Telescope, of the order of 1.5–1.8 arcsec, which will cause the fibre to average over a large fraction of the total galaxy light in most cases and dilute aperture effects. An overabundance of late-type galaxies was detected at redshifts beyond 0.11, which could be attributed to either aperture effects or evolution; however, this will not affect our volume-limited galaxy sample with a maximum redshift of $z_{\max} = 0.114$. Aperture effects are discussed further in Section 6.2.4.

#### 3.2.2 Broad-band colours

More recently it has been possible to obtain broad-band colours for the 2dFGRS galaxies using the same $B_J$ UKST plates as the survey input catalogue (Hambly et al. 2001), but now scanned with the SuperCosmos machine to yield smaller errors of approximately 0.09 mag band$^{-1}$. Similar scans have also been made of the UKST $R_F$ plates. The extinction corrections are from the dust maps of Schlegel, Finkbeiner & Davis (1998) and wavelength-dependent extinction ratios are from Cardelli, Clayton & Mathis (1989). We define rest frame colour

$$(B-R)_0 \equiv B_J - R_F - K(B_J) + K(R_F), \tag{35}$$

where the colour-dependent $K$-corrections are

$$K(B_J) = (-1.63 + 4.53C)\,y + (-4.03 - 2.01C)\,y^2 - \frac{z}{1+(10z)^4}, \tag{36}$$

$$K(R_F) = (-0.08 + 1.45C)\,y + (-2.88 - 0.48C)\,y^2, \tag{37}$$

with $y = z/(1+z)$ and $C = B_J - R_F$. See Cross et al. (2004) for more details. A division at $(B-R)_0 = 1.07$ achieves a similar separation between 'passive' and 'actively star-forming' galaxies to spectral classification of type 1 to types 2–4, giving a total of 77 120 red galaxies and 144 292 blue galaxies.

The distributions of $\eta$ type and colour for the 2dFGRS galaxies are shown in Fig. 1, and the joint distribution is shown in Fig. 2. The correlation between the two properties is clear, together with the distinct bimodality, yet it is obvious that the relationship is not exactly one-to-one. Table 1 gives the respective numbers of each galaxy type in the 2dFGRS catalogue for comparison.

## 4 METHOD: COUNTS-IN-CELLS

A counts-in-cells analysis is employed, which involves splitting the survey region into a lattice of roughly cubical cells and counting the number of galaxies in each cell. The cell dimensions are defined such that all have equal volume $V \equiv L^3$, but with limits to right ascension and declination that form a square on the sky. This angular selection simplifies the treatment of the survey mask, but it means that the cells are not perfect cubes. Over the redshift range involved, this effect is small. The cells are required to fit strictly within the 2dFGRS area, causing some parts of the survey to be unused. Although this restriction, in principle, removes any boundary effects, it means that cells of different sizes sample slightly different areas of the

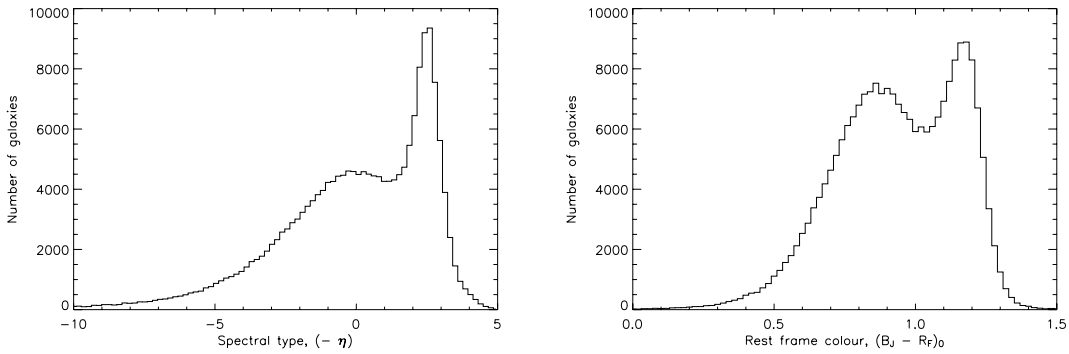


**Figure 1.** The distributions of spectral-type and rest frame colour for all 2dFGRS galaxies. The distinction between passive and actively star-forming galaxies is clear in both distributions. Cuts at $\eta = -1.4$ and $(B - R)_0 = 1.07$ produce the four subgroups with which we work.

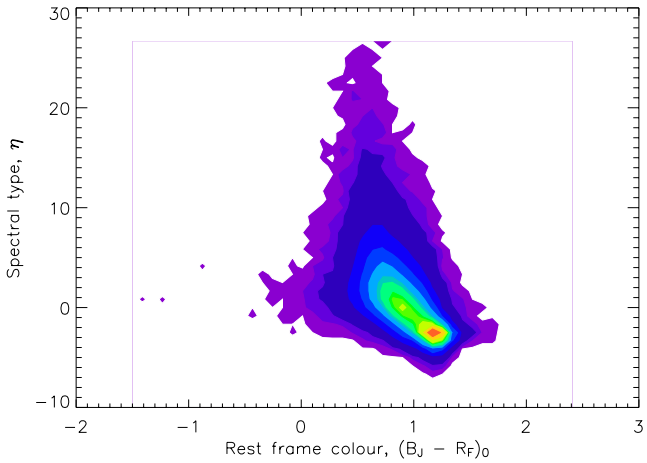


**Figure 2.** The joint distribution of the colour and spectral types for galaxies in the 2dFGRS.

**Table 1.** Numbers of early/late and red/blue galaxies in the 2dFGRS catalogue with good-quality spectra ($Q \geqslant 3$).

| | Red/early | Blue/late | Total |
|---|---|---|---|
| Colour | 77 120 | 144 292 | 221 414 |
| $\eta$ | 74 548 | 118 424 | 192 979 |

Universe. We define our cells with $h = 0.7$ and in what follows all cell lengths are quoted in Mpc, instead of the standard $h^{-1}$ Mpc. In practice, we considered $10 \leqslant L \leqslant 45$ Mpc, giving a total of between 11 423 and 72 cells in the volume-limited survey area after removing low completeness cells. These cell sizes are equivalent in volume to using top-hat smoothing spheres with radii $6.1 \leqslant r \leqslant 27.9\,h^{-1}$ Mpc. Fig. 3 shows an example of how 25-Mpc cells cover the 2dFGRS volume to $z = 0.11$.

As a result of internal holes in the survey and the adaptive tiling algorithm employed, the sampling fraction in the 2dFGRS varies over the sky. Random 2dFGRS catalogues can be created, which include these selection effects by making use of the calculated survey masks. Each cell count is weighted by the fraction of random points found in the same cell in the mock catalogue. The spectral-type analysis introduces extra selection effects, which are quantified by a special mask (see Section 3.1).

An overdensity $\delta$ is calculated for each cell by dividing the observed cell counts $N$ by the expected number for a given cell allowing for completeness, $\bar{N}$:

$$\delta_i = \frac{N_i}{\bar{N}} - 1. \tag{38}$$

This procedure is carried out for both early- and late-type galaxies within each cell. The overall density variance is defined by equation (4).

It is necessary to set a completeness limit to remove excessively undersampled cells, such as those affected by holes in the survey due to stars, or cells at the less observed edges of the survey volume. Although the limits are somewhat arbitrary, it is important they are set correctly as incomplete cells could affect our measurements of scatter in the biasing relation. Fig. 4 shows the completeness distributions for $L = 25$-Mpc cells. It can be seen that for both $\eta$ and colour classification the distribution has a sharp peak of almost complete cells, with a long tail to low completeness and a sharp cut-off at high completeness. The figure also highlights the importance of including the effects of $\eta$ classification on the 2dFGRS mask as the completeness peak and cut-off is noticeably lower for $\eta$ classification than for all galaxies. The lower completeness is reflected in the reduced number of cells available for analysis.

A completeness limit is set for each cell at 70 per cent (or 60 per cent for the larger cells), to include all cells within the high completeness peak. In order to check the effects of completeness on the final results, the models were also fitted to only those cells with a completenesses higher than 80 per cent (70 per cent for the larger cells), and the results were found to be consistent within the errors.

A general concern with a counts-in-cells analysis of observational data is the varying survey selection function over the extent of a cell. For example, a cell containing a cluster of galaxies at its most distant edge, and weighted by the average selection function over its volume, would give a different 'count' to a cell containing a cluster near its inner boundary. Furthermore, with a joint counts-in-cells analysis any relationship between luminosity and galaxy type or colour will cause differing fractions of objects with redshift.

With careful use of type-dependent selection functions such effects can be allowed for (Conway et al. 2004), but in our analysis we use volume-limited samples. This approach avoids these complications at the expense of reducing the number of galaxies in the analysis. The size of the 2dFGRS offers a great advantage over previous observational studies of relative bias, because it provides volume-limited samples large enough to produce reliable measurements.

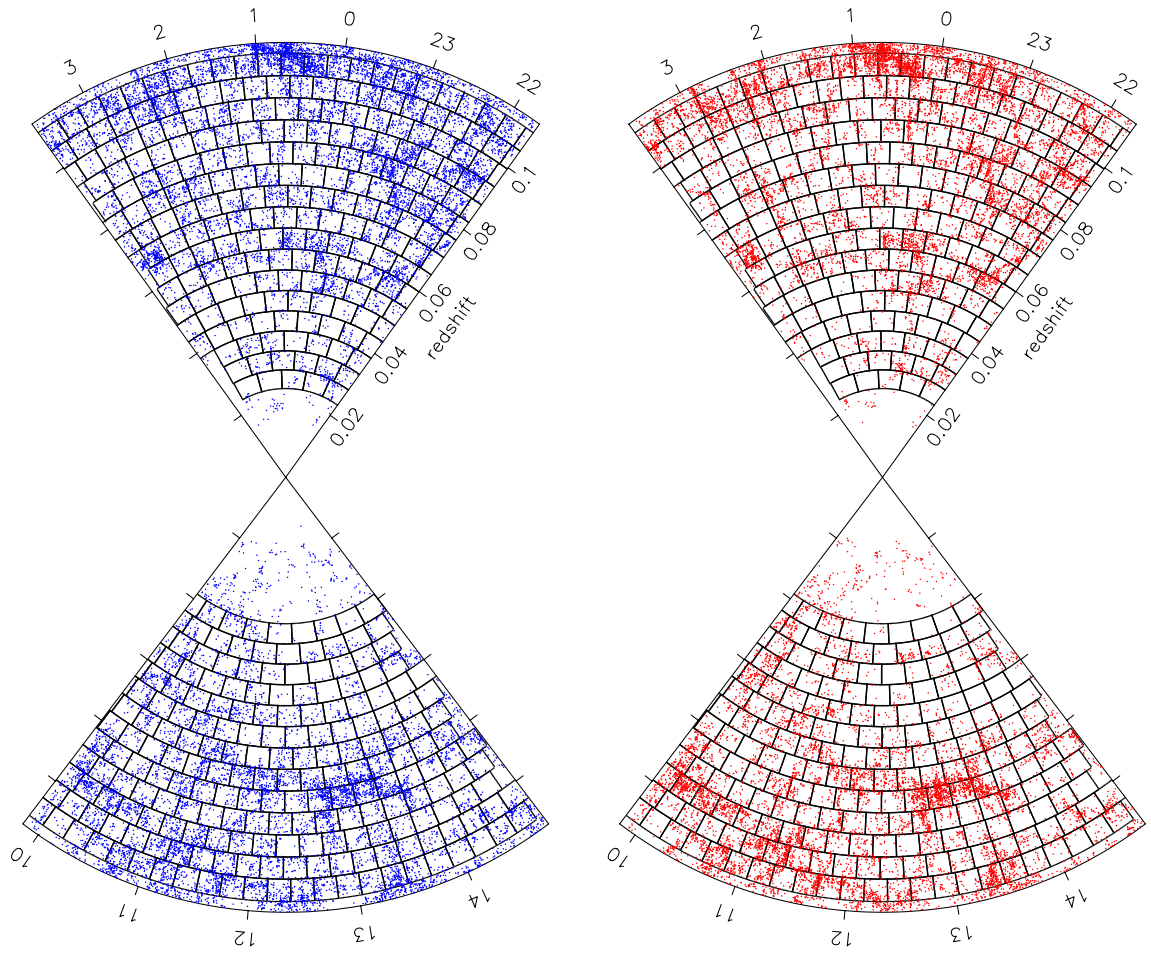


**Figure 3.** Wedge plots of the 2dFGRS volume-limited survey region with $M_{B_J} - 5\log_{10}(h) \leqslant -19.0$. Dots represent late-type galaxies on the left and early-type galaxies on the right (classified by spectral type). The redshift increases from the centre, and right ascension is shown on the horizontal axis, declination is projected on to the plane. Typical cell boundaries of length $L = 25$ Mpc are overplotted.

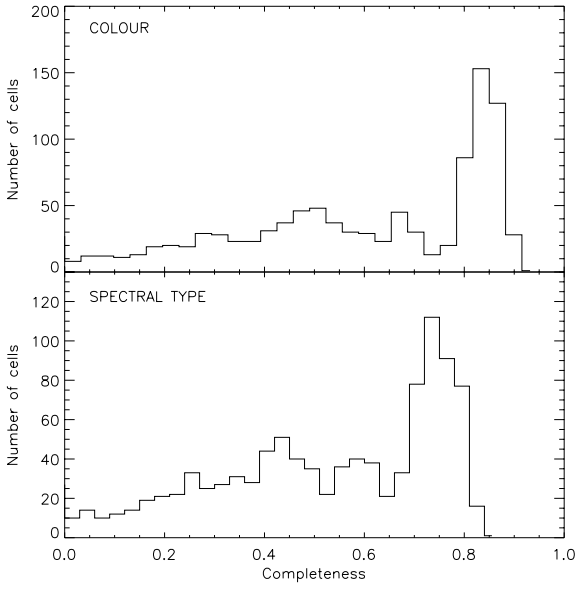


**Figure 4.** Histograms showing the completeness distribution of 25-Mpc cells using the standard 2dFGRS mask (top) and including the effects of $\eta$ classification (bottom).

## 5 PARAMETER FITTING

### 5.1 A maximum-likelihood approach

Once we have chosen a model, a maximum-likelihood method is used to fit the free parameters of the model to the data. Denoting the number of early- (late-) type galaxies within cell $i$ as $N_{\mathrm{E},i}$ ($N_{\mathrm{L},i}$), the likelihood of finding a cell containing $N_{\mathrm{E}}$ early-type and $N_{\mathrm{L}}$ late-type galaxies given a model with free parameters $\boldsymbol{\alpha}$, is defined as

$$L_i(N_{\mathrm{E},i}, N_{\mathrm{L},i}; \boldsymbol{\alpha}) = P(N_{\mathrm{E},i}, N_{\mathrm{L},i}|\boldsymbol{\alpha}) \tag{39}$$

and the total likelihood for all cells is then

$$L = \prod L_i. \tag{40}$$

The likelihood can be maximized with respect to the free parameters $\boldsymbol{\alpha}$ to find the best-fitting values $\hat{\boldsymbol{\alpha}}$ for the model given the data set. In practice it is easier to minimize the function

$$\mathcal{L} \equiv -\sum_i \ln L_i. \tag{41}$$

Note that this definition of $\mathcal{L}$ differs by a factor of 2 compared with Conway et al. (2004).

The models in Section 2 contain two or three free parameters: $\sigma_{\rm E}$ and/or $\sigma_{\rm L}$ from the one-point PDFs, and $b$ or $r_{\rm LN}$ from the conditional probability function. These parameters were fitted simultaneously to the data using a downhill simplex method (Press et al. 1992).

### 5.2 Error estimation

As it is not possible to derive analytic solutions to the sampling distribution of our maximum-likelihood estimators $\hat{\boldsymbol{\alpha}}$, the standard error on our parameters must be estimated directly from the likelihood function using Bayes' theorem and assuming a uniform prior on $\boldsymbol{\alpha}$:

$$P(\boldsymbol{\alpha}|\boldsymbol{x}) \propto L(\boldsymbol{x};\boldsymbol{\alpha}), \tag{42}$$

where $\boldsymbol{\alpha}$ again denotes the model parameters, $\boldsymbol{x}$ the data and $P$ the probability.

For a single free parameter, the upper and lower limits on $\alpha$ are found from

$$P(\alpha_- \leqslant \alpha < \alpha_+|\boldsymbol{x}) = \frac{\int_{\alpha_-}^{\alpha_+} L(\boldsymbol{x};\alpha)\,{\rm d}\alpha}{\int_{-\infty}^{\infty} L(\boldsymbol{x};\alpha)\,{\rm d}\alpha}. \tag{43}$$

If it can be assumed that the likelihood function is reasonably approximated by a Gaussian, $1\sigma$ errors on the parameter can be estimated. For multiparameter models it is necessary to quantify any possible degeneracy between errors. If the multidimensional likelihood function can be approximated by a multivariate Gaussian distribution, individual errors and correlations between the parameters can be found.

A second method of error estimation involves creating many mock data sets from the fitted model probability distributions themselves. These data sets are made through Monte Carlo techniques, and designed to closely reproduce the true data in size. On applying the above likelihood techniques to these mock data sets, the best fit and true parameters can be compared with estimate the errors. The advantage of this method is that no assumptions need to be made concerning the shape of the likelihood function. The disadvantage is that we are assuming that the model is a correct representation of the data, as the errors strictly apply only to the model not the data. By increasing the size of the mock data sets, this method can also be used to check for any bias inherent in the fitting method. This process was carried out for each model in this paper, finding the parameter estimations to be unbiased.

In all cases, we will make the assumption that the density fluctuations in each cell can be treated as being independent. This is clearly not true in detail, as the existence of modes with wavelength $\gtrsim L$ will cause a correlation between nearby cells. This was considered by Broadhurst, Taylor & Peacock (1995), who showed that the correlation coefficient was low even for adjacent cells: $r \simeq 0.2$. As we shall see, it is $(1 - r^2)^{1/2}$ that matters for joint distributions, and so the failure of independence is negligible in practice.

### 5.3 Model comparison

Once we have found the best-fitting parameters for each of our three models, we would like to know the goodness-of-fit of the models and the significance of any differences between the fits. We approach this using two different methods.

#### *5.3.1 Likelihood ratios*

To test the significance of one model against another model we use the likelihood ratio test. In its simplest form we define the maximum-likelihood ratio for hypothesis $H_0$ versus $H_1$

$$\lambda = \frac{L(\boldsymbol{x}|H_0)}{L(\boldsymbol{x}|H_1)}, \tag{44}$$

where $\boldsymbol{x}$ is the data and $L$ represents the *maximum* likelihood value. This will be especially valuable in assessing the evidence for stochasticity, where we will compare a model of perfect correlation with one where $r_{\rm LN} \neq 1$ is allowed, effectively introducing an extra parameter. The key question is how large a boost in likelihood is expected from the introduction of an extra parameter, and this was considered by Liddle (2004). He advocates the *Bayesian information criterion*, defined as

$$B = -2\ln L + p\ln N, \tag{45}$$

where $p$ is the number of parameters and $N$ is the number of data points. This measure of information effectively says that going from a satisfactory model with $p$ parameters to one that overfits with $p + 1$ parameters would be expected to increase $\ln L$ by $0.5\ln N$. Therefore, in order to achieve evidence in favour of the increase to $p + 1$ at the usual 5 per cent threshold, we require

$$\Delta\ln L = -\ln 0.05 + 0.5\ln N, \tag{46}$$

which is between 5 and 8 for the number of cells considered here. An unequivocal detection of stochasticity thus apparently requires a likelihood ratio of between $r_{\rm LN} \neq 1$ and the best $r_{\rm LN} = 1$ model in excess of $\lambda \simeq \exp(5)$ to $\exp(8)$.

Monte Carlo simulations may be used to check the validity of this analytic method. This is computationally expensive, so only an upper limit may be set on the significance of an observed likelihood ratio. We create 40 mock data sets following a power-law bias model convolved with a Poisson distribution, defining the mean cell counts, number of cells, one-point PDF fit parameter $\sigma_{\rm E}$ and model parameter $b_{\rm pow}$ to emulate a range of data sets. To these we fit both power-law and bivariate lognormal models with the usual maximum-likelihood fitting procedure. This allows us to assess the largest likelihood ratio that should arise by chance if the true model is, in fact, a perfect power-law bias. The results suggest a substantially smaller critical value is required than equation (46), closer to $\Delta\ln L = 1$ to reject the model at the 95 per cent confidence limit. It therefore appears that the assumptions used to derive the Bayesian information criterion do not apply to this problem.

#### *5.3.2 Kolmogorov–Smirnov test*

Although the likelihood ratio test can eliminate one model in favour of another, it cannot tell us how well the preferred model fits the data. A Kolmogorov–Smirnov (KS) test can be used to test for a difference between an observed and modelled cumulative probability distribution. This test provides the probability that the data are drawn from the model probability distribution, with a low probability representing a poor fit. A resulting probability above approximately 0.1 is generally accepted as a reasonable fit, as the KS test is unable to rule out the model being the true underlying distribution at greater than 90 per cent confidence. Strictly the test becomes invalid once the data has been used to fix any free parameters of the model, as in this method (Lupton 1993). However, as long as the number of data points is much greater than the number of free parameters any effects should be small.

To transform our bivariate distribution to a one-dimensional variable on which we can perform the standard KS test, we create integrated probability distributions from both the model and the data, integrating within constant model probability contours centred on

**Table 2.** Completeness limits, total number of cells and average cell counts for each data set after corrections for completeness have been applied. Note that numbers do not scale exactly as $L^3$ due to edge effects.

| | Cell size (Mpc) | Compl. | No cells | $\langle N_{\rm E}\rangle$ | $\langle N_{\rm L}\rangle$ |
|---|---|---|---|---|---|
| Colour | 10 | 0.7 | 11 423 | 1.5 | 1.8 |
| | 15 | 0.7 | 3019 | 5.1 | 6.3 |
| | 20 | 0.7 | 1104 | 12.1 | 14.6 |
| | 25 | 0.7 | 484 | 25.6 | 30.1 |
| | 30 | 0.7 | 234 | 41.3 | 48.6 |
| | 35 | 0.6 | 169 | 57.4 | 70.7 |
| | 40 | 0.6 | 115 | 88.3 | 105.4 |
| | 45 | 0.6 | 72 | 125.9 | 149.2 |
| $\eta$ | 10 | 0.7 | 9668 | 1.9 | 1.9 |
| | 15 | 0.7 | 2567 | 6.5 | 6.3 |
| | 20 | 0.7 | 930 | 15.3 | 14.7 |
| | 25 | 0.7 | 404 | 32.1 | 30.2 |
| | 30 | 0.7 | 187 | 54.2 | 49.6 |
| | 35 | 0.6 | 115 | 71.8 | 71.4 |
| | 40 | 0.6 | 74 | 106.4 | 108.7 |

the position of maximum probability. This gives cumulative probability distributions for model and data from which the KS probability (that the data follow the same underlying distribution as the model) can be derived.

The KS test has been generalized to bivariate analyses by Peacock (1983) and Fasano & Franceschini (1987). However, this two-dimensional KS test was found to lack power compared with the previous method for the present application.

## 6 RESULTS

Table 2 summarizes the details of each of our samples, including the average count per cell and the completeness limit. The following sections look in detail at the univariate and bivariate model fits to each data set. In Section 6.2.3 we investigate the scale dependence of non-linearity and stochasticity in the 2dFGRS. In Sections 6.3 and 6.4 we discuss the origin of the stochasticity and perform some consistency checks on the results.

### 6.1 One-point probability distributions

Fig. 5 shows the bivariate distributions of cell counts, together with the one-point distribution functions for a range of cell sizes. Before we consider the bivariate distributions further, we look in detail at the individual lognormal fits to these one-point distributions. We fit a lognormal distribution convolved with Poisson noise to the early and late number counts individually, using the method described in the previous sections. The best-fitting lognormal models are shown overplotted in the figure. It can be seen that on large scales the lognormal model alone fits the data well, but on small scales the deviation due to discreteness is substantial. For this reason it is important to account for shot noise in the fitting procedure.

In order to assess the Poisson-sampled lognormal model quantitatively, we create many Monte Carlo cells with best-fitting parameters and expected number counts to match each data set. We allow for completeness effects by randomly assigning each cell a completeness value from a Gaussian distribution of width and mean equal to those of the data set. Fig. 6 shows the distributions of early-type galaxies and Monte Carlo cells with matching parameters. Overplotted are the best-fitting lognormal model (dashed line) and a lognormal curve with variance derived directly from the data (dotted line, see Section 6.5).

We can now compare our MC data with our true data through a KS test. For large cells ($\geqslant 25$ Mpc) we find KS probabilities in excess of 0.8, but as cell size decreases the KS probabilities decrease. On the smallest scales of 10 Mpc we obtain KS probabilities of $\sim 10^{-9}$ and it is this poor model fit that causes the lognormal model to overestimate variances in comparison with direct methods. The figure shows there to be an excess of data cells with moderate overdensities compared with the best-fitting lognormal model, particularly on the smallest scales. In very underdense regions the figure shows the dotted (direct variance) curve to lie below the dashed (fitted) curve. This results in an underprediction of the number of cells containing zero or one galaxy when using the direct variance method (as discussed in more detail by Conway et al. 2004).

#### 6.1.1 *Failures of the Poisson-sampled lognormal distribution*

The discrepancy between the observed and predicted distributions of cell counts shows that at least one of our two assumptions concerning the galaxy field is incorrect. The lognormal distribution is simply a convenient functional form that has been shown to fit galaxy distributions from previous surveys (e.g. Hamilton 1985; Kofman et al. 1994) and *N*-body matter distributions successfully (Kofman et al. (1994); Kayo, Taruya & Suto (2001) and references therein). Deviations from this simple model are evident in detailed numerical simulations (e.g. Bernardeau & Kofman 1995), and at some level at least we would expect to see such deviations in our data. Various alternative distributions have been suggested in the literature such as the skewed lognormal, negative binomial or Edgeworth expansions (see also Sheth, Mo & Saslaw 1994; Valageas & Munshi 2004).

However, the fact that the model fails in detail on scales at which the shot noise dominates the distribution in underdense regions suggests that the *Poisson sampling hypothesis* is at least partly to blame. By attempting to fit the model to these underdense cells, the variance is increased and the moderately overdense regions are no longer well fitted. On small scales the majority of cells contain zero or one galaxy, hence the preference of the model to fit these cells and not those containing more galaxies.

We hope to explore more complex models for the count distribution elsewhere. For the present application, there are two points to make. The first is that cells of side approximately 10 Mpc are the smallest that can sensibly be discussed with this approach; reducing the cell size would lead to distributions that are dominated by discreteness effects. More importantly, it should be stressed that analytic models of this sort are not really physical. In the end, what matters is whether the 2dFGRS data match the predictions of a proper calculation of galaxy formation. We carry out such a comparison at the end of the paper, and the results of a Poisson-sampled lognormal fit are a convenient statistic to use for this purpose. Provided true data and mock data are treated identically, small imprecisions in the function used for the fit are irrelevant.

### 6.2 Joint distributions and biasing models

Each of the three models in Section 2 is fitted to the data sets described in Table 2. Best-fitting parameters for the models are estimated simultaneously through the maximum-likelihood method of Section 5.1. Table 3 shows the best-fitting parameter values for the two deterministic models, together with log-likelihood differences between the model and the bivariate lognormal model. The

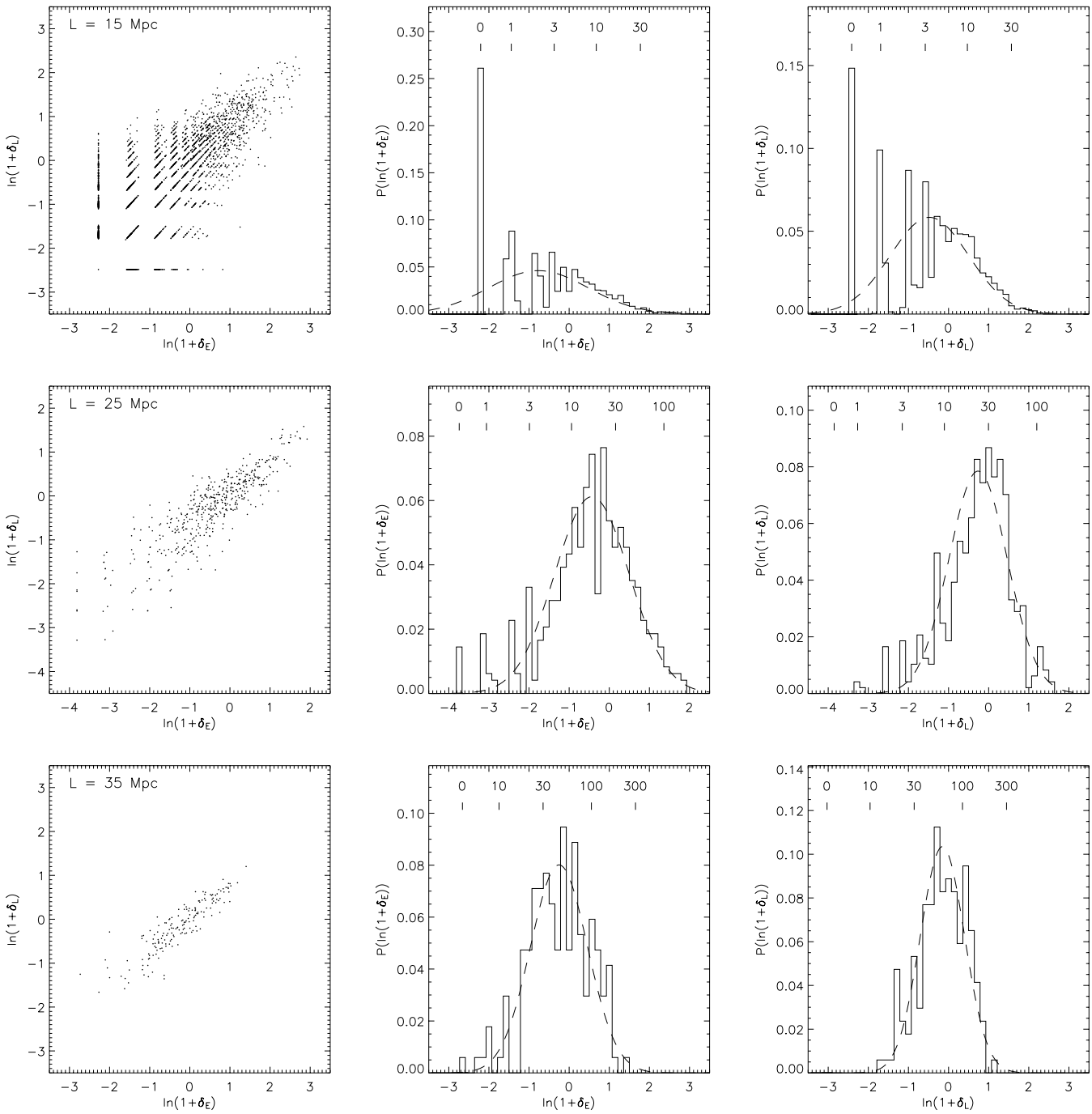


**Figure 5.** On the left, the bivariate counts-in-cells distributions with early- and late-type galaxies are shown classified by colour. The points mark density values of individual cells, and from top to bottom $L = 15$-, 25- and 35-Mpc cells are shown. 1D projections of the distributions are shown for early types (centre) and late types (right), to which a Poisson-sampled lognormal model has been fitted. The best–fitting lognormal curves are overplotted (dashed line). As a result of the logarithmic axes, a bin for cells containing zero galaxies has been artificially positioned on the horizontal axis. Note the discreteness of the galaxy counts: the actual number of galaxies contained in the cells is indicated by the numbers over the 1D distributions. Further note the survey completeness effects on smaller counts per cell, causing the spread of points around the mean value. Correcting zero counts for completeness is non-trivial and is not included in this analysis, hence there is no spread of these points.

values of $b_{\rm pow}$ and $b_{\rm lin}$ clearly show how early-type galaxies are more clustered than late-type galaxies, as is well known. Table 4 gives the best-fitting parameter values and errors for the stochastic bias model. The quoted errors are determined through multivariate Gaussian fits to the likelihood surface, which were found to agree well with Monte Carlo error estimates.

Fig. 7 shows the joint probability distribution of the data for $L =$ 20-Mpc cells, together with Monte Carlo realizations of the best-fitting linear, power-law and bivariate lognormal models for comparison. The Monte Carlo realizations include completeness effects by randomly selecting a completeness value for each cell from a Gaussian with mean and width equal to that of the true distribution of cell

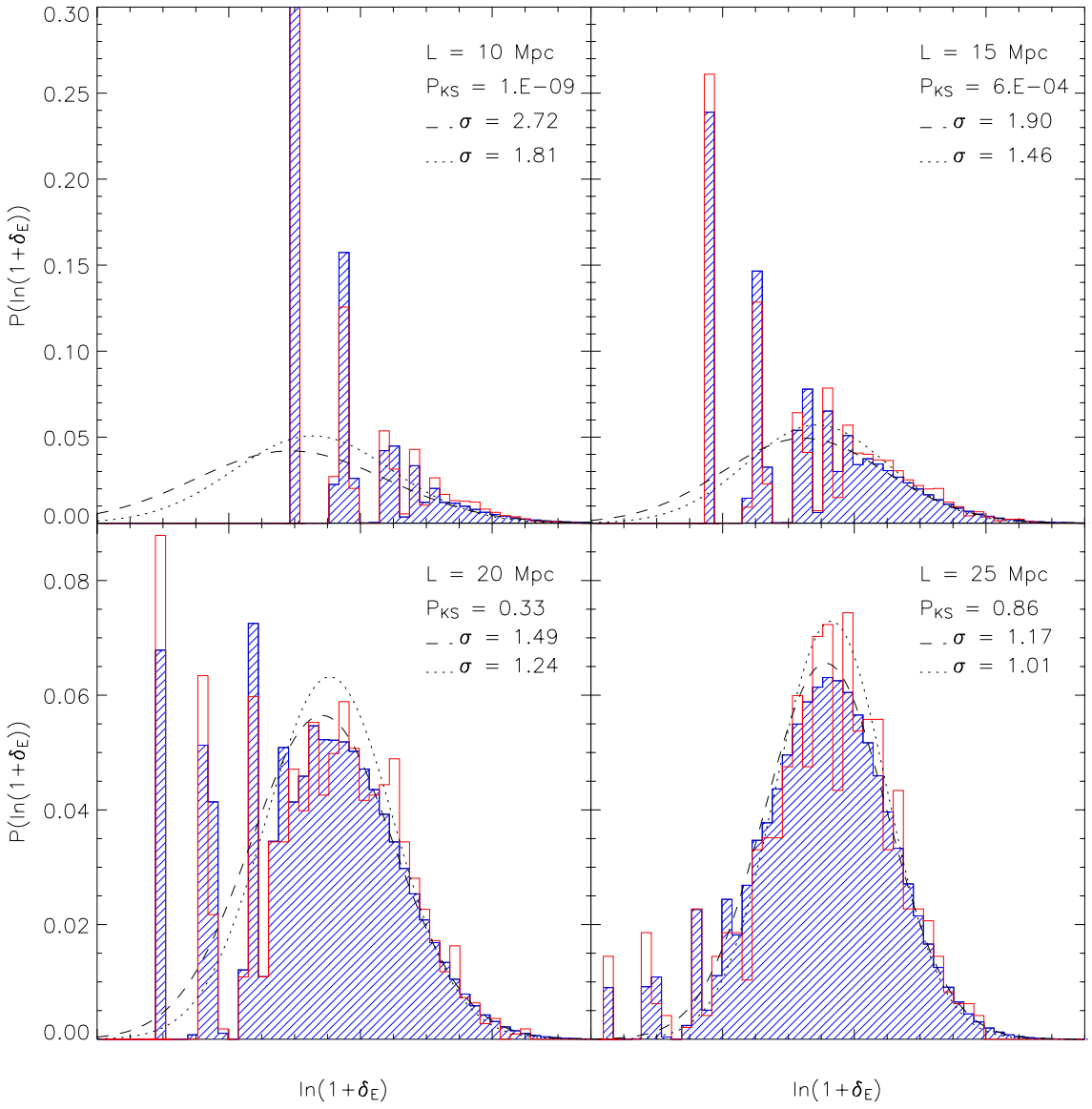


**Figure 6.** The univariate distributions of early-type galaxies for $L = 10$-, 15-, 20- and 25-Mpc cells [empty, red (black) histogram], together with the distribution of Monte Carlo cells [hatched, blue (grey) histogram] with parameters equal to those obtained from fitting a Poisson-sampled lognormal distribution to the data cells. Completeness effects are modelled as a Gaussian with variance equal to that found in the data set. The dashed curve shows this best-fitting lognormal distribution; the dotted curve shows the lognormal curve with variance equal to that measured directly from the data. Both variances are given in the upper right-hand side, together with KS probabilities that the Monte Carlo cells are drawn from the same distribution as the data cells. As a result of the logarithmic axes, a bin for cells containing zero galaxies has been artificially placed on the horizontal axis. For $L = 10$ Mpc approximately 60 per cent of the cells contain zero early-type galaxies, and the vertical axis has been truncated to allow a better view of the remaining bins. Note the discreteness of the data and simulations, as discussed in the caption to Fig. 5.

completeness. This scale is chosen to illustrate all the properties of the data and the figures clearly show the shot noise in underdense regions, together with the effects of survey incompleteness. By eye we can see the differences between the linear and power-law bias models, and the effect of stochasticity in the bivariate lognormal model. The best-fitting linear model has a mean bias closer to unity at high density, and cannot fit the non-linearity seen in the data. The power-law model corrects for this, but the scatter about the mean is insufficient to match the data. The stochasticity introduced by the bivariate lognormal model is evident and is matched well by the data. The likelihood ratios shown in Table 3 quantify the differences and show that on all scales the bivariate lognormal model gives significantly better fits compared with deterministic biasing models.

We now repeat this analysis, splitting galaxies by spectral type $\eta$, rather than by colour. The colour split allows a larger sample of cells to be included in the analysis, but Fig. 2 shows that a division by spectral type does not always select the same galaxies as a colour split, so it is interesting to see how the results compare. The second section of Tables 2–4 give details of the data sets and results of the model fits to cells with galaxies classified by $\eta$. The joint distributions for 20-Mpc cells are shown in Fig. 8.

Comparing the results for 20-Mpc cells, the results for colour and $\eta$ are generally similar. The likelihood ratios again favour the

**Table 3.** The best-fitting deterministic biasing models parameters to each data set. The level of non-linearity given the model is given by $\tilde{b}/\hat{b}$, which is unity by definition for the linear model. The penultimate column shows the log-likelihood differences between the best-fitting linear or power-law models and the bivariate lognormal model. A positive value indicates that the bivariate lognormal model is a better fit to the data. The final column shows how many cells must be removed to reduce the power-law likelihood ratio to $\sim \exp(1)$ (Section 6.3).

| | Cell size | Model | $\omega_E$ | $b_{lin}$ or $b_{pow}$ | $\hat{b}$ | $b_{var}$ | $r_{lin}$ | $\tilde{b}/\hat{b}$ | $\mathcal{L}-\mathcal{L}^{LN}$ | Outliers |
|---|---|---|---|---|---|---|---|---|---|---|
| Colour | 10 | Linear | 1.41 | 0.93 | 0.93 | 0.93 | 1.00 | 1.000 | 291.1 | |
| | | Power-law | 1.51 | 0.78 | 0.56 | 0.58 | 0.96 | 1.044 | 75.6 | 159 |
| | 15 | Linear | 1.23 | 0.91 | 0.91 | 0.91 | 1.00 | 1.000 | 185.0 | |
| | | Power-law | 1.26 | 0.77 | 0.61 | 0.63 | 0.97 | 1.030 | 55.5 | 86 |
| | 20 | Linear | 1.04 | 0.91 | 0.91 | 0.91 | 1.00 | 1.000 | 133.6 | |
| | | Power-law | 1.10 | 0.76 | 0.63 | 0.65 | 0.98 | 1.024 | 37.9 | 40 |
| | 25 | Linear | 0.92 | 0.91 | 0.91 | 0.91 | 1.00 | 1.000 | 87.1 | |
| | | Power-law | 0.94 | 0.76 | 0.67 | 0.68 | 0.98 | 1.016 | 41.3 | 34 |
| | 30 | Linear | 0.80 | 0.92 | 0.92 | 0.92 | 1.00 | 1.000 | 39.3 | |
| | | Power-law | 0.77 | 0.77 | 0.72 | 0.72 | 0.99 | 1.009 | 22.4 | 18 |
| | 35 | Linear | 0.73 | 0.92 | 0.92 | 0.92 | 1.00 | 1.000 | 16.0 | |
| | | Power-law | 0.70 | 0.76 | 0.71 | 0.72 | 0.99 | 1.008 | 6.6 | 6 |
| | 40 | Linear | 0.67 | 0.95 | 0.95 | 0.95 | 1.00 | 1.000 | 18.1 | |
| | | Power-law | 0.68 | 0.81 | 0.77 | 0.78 | 1.00 | 1.005 | 12.4 | 8 |
| | 45 | Linear | 0.59 | 0.97 | 0.97 | 0.97 | 1.00 | 1.000 | 12.9 | |
| | | Power-law | 0.59 | 0.86 | 0.83 | 0.83 | 1.00 | 1.002 | 10.4 | 1 |
| $\eta$ | 10 | Linear | 1.43 | 0.92 | 0.92 | 0.92 | 1.00 | 1.000 | 238.1 | |
| | | Power-law | 1.51 | 0.77 | 0.55 | 0.58 | 0.96 | 1.046 | 49.8 | 110 |
| | 15 | Linear | 1.22 | 0.91 | 0.91 | 0.91 | 1.00 | 1.000 | 131.1 | |
| | | Power-law | 1.24 | 0.77 | 0.64 | 0.65 | 0.97 | 1.026 | 35.6 | 58 |
| | 20 | Linear | 1.03 | 0.90 | 0.90 | 0.90 | 1.00 | 1.000 | 91.9 | |
| | | Power-law | 1.07 | 0.75 | 0.67 | 0.68 | 0.98 | 1.019 | 17.9 | 25 |
| | 25 | Linear | 0.93 | 0.90 | 0.90 | 0.90 | 1.00 | 1.000 | 70.3 | |
| | | Power-law | 0.96 | 0.74 | 0.65 | 0.67 | 0.98 | 1.018 | 23.7 | 21 |
| | 30 | Linear | 0.81 | 0.90 | 0.90 | 0.90 | 1.00 | 1.000 | 28.4 | |
| | | Power-law | 0.81 | 0.73 | 0.67 | 0.68 | 0.99 | 1.013 | 13.3 | 9 |
| | 35 | Linear | 0.74 | 0.90 | 0.90 | 0.90 | 1.00 | 1.000 | 13.6 | |
| | | Power-law | 0.75 | 0.69 | 0.65 | 0.66 | 0.99 | 1.013 | 4.3 | 2 |
| | 40 | Linear | 0.64 | 0.95 | 0.95 | 0.95 | 1.00 | 1.000 | 4.8 | |
| | | Power-law | 0.65 | 0.83 | 0.82 | 0.82 | 1.00 | 1.003 | 1.2 | 0 |

**Table 4.** The best-fitting bivariate lognormal model parameters to each data set. Errors are shown, derived from Gaussian fits to the parameter likelihood surface. $\Delta(r_{LN})$ is derived from propagation of $\Delta[(1-r_{LN}{}^2)^{1/2}]$. The remaining columns give the average biasing parameters. The Appendix gives the analytic solutions for each parameter in the case of the bivariate lognormal model. The final two parameters measure the non-linearity and stochasticity of the model (equations 13 and 14).

| | Cell size | $\omega_E$ | $\Delta(\omega_E)$ | $\omega_L$ | $\Delta(\omega_L)$ | $r_{LN}$ | $\Delta(r_{LN})$ | $\sigma_E$ | $\sigma_L$ | $r_{lin}$ | $\hat{b}$ | $b_{var}$ | $\tilde{b}/\hat{b}$ | $\sigma_b/\hat{b}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Colour | 10 | 1.52 | 0.01 | 1.20 | 0.01 | 0.958 | 0.004 | 3.01 | 1.80 | 0.88 | 0.52 | 0.55 | 1.054 | 0.44 |
| | 15 | 1.26 | 0.02 | 0.99 | 0.02 | 0.966 | 0.004 | 1.99 | 1.29 | 0.92 | 0.60 | 0.62 | 1.033 | 0.35 |
| | 20 | 1.10 | 0.02 | 0.85 | 0.02 | 0.969 | 0.005 | 1.54 | 1.02 | 0.93 | 0.62 | 0.64 | 1.026 | 0.31 |
| | 25 | 0.95 | 0.02 | 0.73 | 0.02 | 0.959 | 0.007 | 1.21 | 0.84 | 0.93 | 0.64 | 0.66 | 1.020 | 0.34 |
| | 30 | 0.78 | 0.03 | 0.61 | 0.03 | 0.962 | 0.009 | 0.92 | 0.68 | 0.94 | 0.69 | 0.70 | 1.011 | 0.31 |
| | 35 | 0.71 | 0.04 | 0.54 | 0.03 | 0.976 | 0.009 | 0.81 | 0.58 | 0.96 | 0.70 | 0.70 | 1.009 | 0.24 |
| | 40 | 0.68 | 0.05 | 0.55 | 0.04 | 0.971 | 0.009 | 0.76 | 0.60 | 0.96 | 0.75 | 0.75 | 1.006 | 0.27 |
| | 45 | 0.60 | 0.04 | 0.51 | 0.04 | 0.970 | 0.010 | 0.66 | 0.55 | 0.96 | 0.80 | 0.80 | 1.003 | 0.27 |
| $\eta$ | 10 | 1.51 | 0.02 | 1.18 | 0.01 | 0.963 | 0.005 | 2.95 | 1.75 | 0.89 | 0.53 | 0.56 | 1.052 | 0.40 |
| | 15 | 1.23 | 0.02 | 0.98 | 0.02 | 0.966 | 0.005 | 1.89 | 1.27 | 0.92 | 0.62 | 0.64 | 1.029 | 0.34 |
| | 20 | 1.07 | 0.02 | 0.82 | 0.02 | 0.976 | 0.005 | 1.47 | 0.98 | 0.94 | 0.63 | 0.64 | 1.025 | 0.27 |
| | 25 | 0.92 | 0.02 | 0.70 | 0.02 | 0.966 | 0.007 | 1.16 | 0.79 | 0.94 | 0.64 | 0.65 | 1.019 | 0.31 |
| | 30 | 0.81 | 0.05 | 0.60 | 0.04 | 0.965 | 0.010 | 0.97 | 0.66 | 0.94 | 0.64 | 0.65 | 1.016 | 0.30 |
| | 35 | 0.73 | 0.04 | 0.51 | 0.03 | 0.980 | 0.009 | 0.84 | 0.55 | 0.96 | 0.63 | 0.64 | 1.015 | 0.22 |
| | 40 | 0.66 | 0.04 | 0.54 | 0.04 | 0.988 | 0.008 | 0.73 | 0.58 | 0.98 | 0.78 | 0.79 | 1.004 | 0.17 |

bivariate lognormal model over our two deterministic biasing models, and suggest a slightly smaller difference between the stochastic and deterministic models than is found in the colour data set. This is verified by the smaller stochasticity found in the best-fitting bivariate lognormal model at all scales. Unlike for the colour data sets, the power-law bias is only marginally inconsistent with the data on the largest scales studied here. This difference between colour and $\eta$ type results may reflect a physical difference

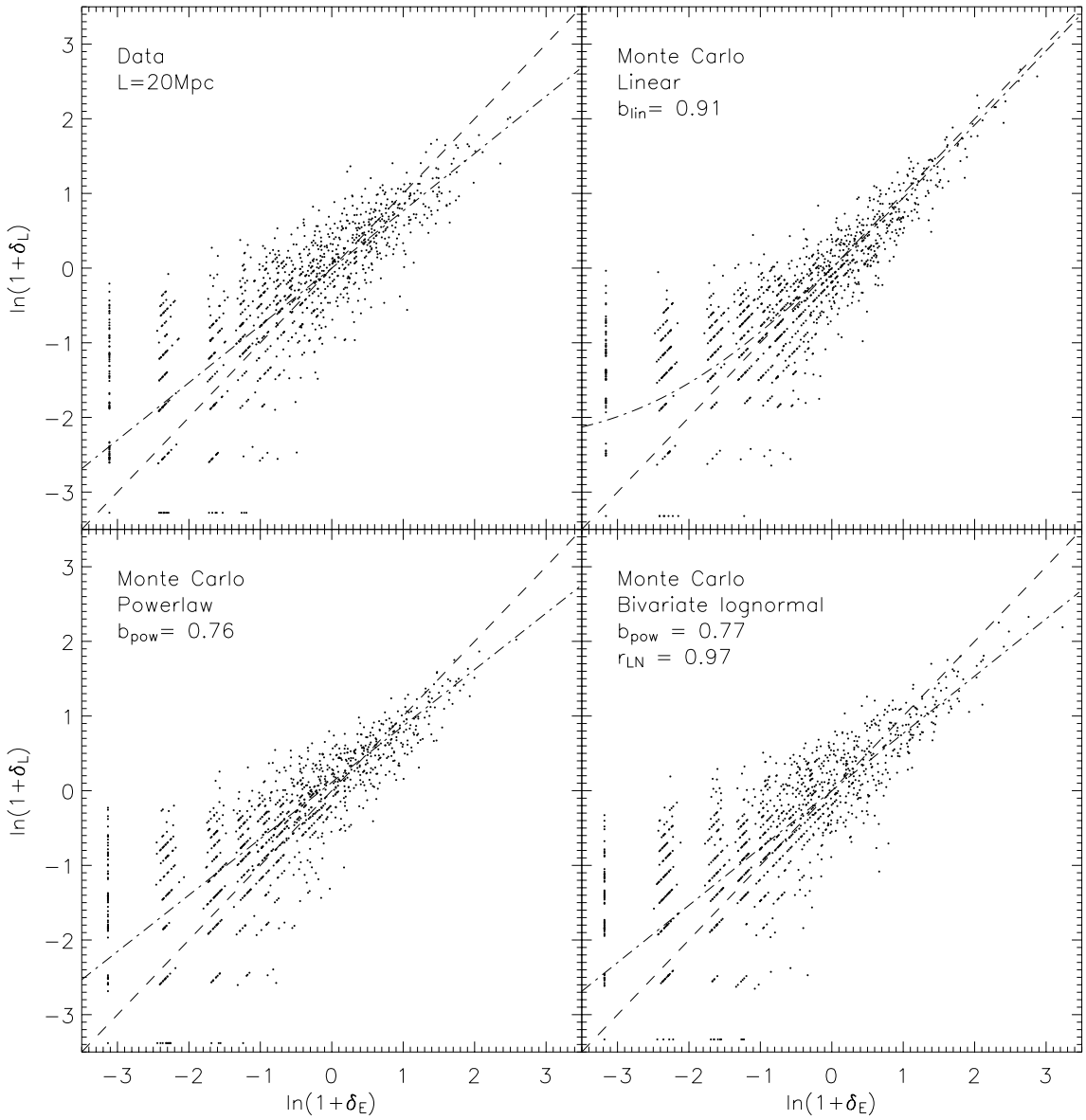


**Figure 7.** On the top left, the bivariate counts-in-cells distributions for 20-Mpc length cells, with early- and late-type galaxies classified by colour. The points mark density values of individual cells. The other three panels show Monte Carlo realizations of the best-fitting linear, power-law and bivariate lognormal models. The realizations are created to match the data as far as possible, with equal cell numbers and average number counts. Cell completeness is included by assuming the distribution of cell completeness to be a Gaussian of mean and width equal to that of the data cells. In each panel the dashed line shows the $b = 1.0$ case, and the dash-dot line shows the mean biasing of each model (for the top left-hand plot, the dash-dot line shows the mean biasing of the best-fitting bivariate lognormal model). Poisson sampling of the galaxies is assumed in all cases. Note that for all but $b = 1.0$, the linear bias appears as a curve on the log–log plots. As a result of the logarithmic axes, cells containing zero early- or late-type galaxies have been artificially positioned.

in the relative biasing relations, but firm conclusions are not yet possible.

### *6.2.1 The goodness-of-fit statistics*

Table 3 shows the log-likelihood differences ($-\ln\lambda$) of the parameter fits, taking the bivariate lognormal model as our null hypothesis. In all cases the linear model provides a worse fit to the data than the power-law model, and the power-law model provides a worse fit than the bivariate lognormal model. This latter statement may, however, be due to the addition of an extra free parameter to the model (Liddle 2004). To establish the significance of the difference between the power-law and bivariate lognormal model we can make use of the theory given in Section 5.3. For example, for $L =$ 20-Mpc cells and assuming the Bayesian Information Criterion (equation 46), we would require a likelihood ratio in excess of exp(6.5) to claim that the bivariate lognormal model provides a significantly better fit than the power-law model with $r_{\rm LN} = 1$. However, as stated earlier, Monte Carlo simulations of the power-law model show that a less stringent likelihood ratio in excess of $\simeq$ exp(1) is all that is required. We measure a likelihood ratio of exp(38) for the data set with galaxies classified by colour, and exp(18) for $\eta$ classification, a highly significant result in both cases.

Although the likelihood ratios favour of the bivariate lognormal distribution over the two deterministic models, they do not tell us how well the best-fitting distribution matches the data. For this we

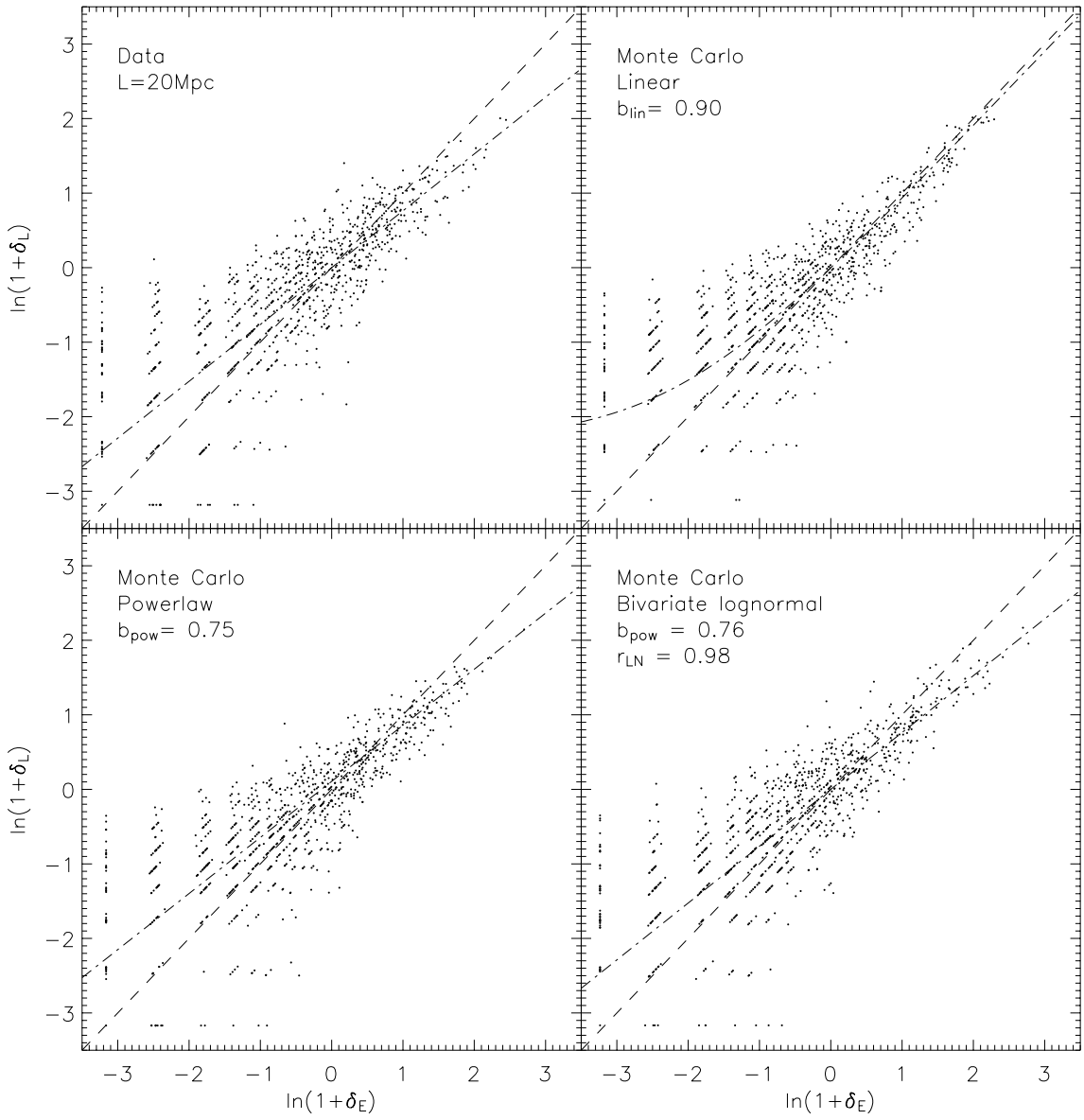


**Figure 8.** Same as in Fig. 7, with galaxies classified by $\eta$ type.

turn to the KS statistic described in Section 5.3.2 On scales $L \geqslant 15$ Mpc our KS statistic accepts the model with a probability greater than 0.5. On the smallest scales studied here we find that this probability decreases, in line with the trend for the univariate lognormal distribution (Section 6.1).

### 6.2.2 *Stochasticity and non-linearity*

In order to quantify the non-linearity and stochasticity of the joint distribution of early- and late-type galaxies we assume that the bivariate lognormal model is an accurate representation of the data. In our analysis the log-density correlation coefficient $r_{\rm LN}$ provides a complete measure of the stochasticity; to aid comparison with other work we compute the mean biasing, its non-linearity and the average biasing scatter of equations (8)–(12). For clarity we concentrate briefly on the results for 20-Mpc cells, shown in the third line of Table 4. These indicate that, whilst the non-linearity (equation 13) is only 1.03, the stochasticity (equation 14) is 0.31. This high stochasticity is reflected in the deviation of $r_{\rm LN}$ from unity, and the low linear correlation coefficient of $r_{\rm lin} = 0.93$. It is important to note that these statistics account for Poisson noise, as the models were convolved with a Poisson distribution before being fitted to the data.

Table 3 shows for comparison some biasing statistics for our two deterministic models. It can be seen that a similar non-linearity is measured by the power-law model, whilst the linear correlation coefficient remains close to unity, reflecting the inability of the model to measure stochasticity. The best-fitting linear bias model has a mean biasing parameter closer to unity, indicating that by assuming this model previous studies may have underestimated the magnitude of relative biasing.

It may be considered surprising that the correlation parameter $r_{\rm LN}$ can be measured so precisely that $r_{\rm LN} = 0.97$ can be clearly distinguished from $r_{\rm LN} = 1$. The reason for this can be seen by examining the expression for the bivariate lognormal distribution (equation 33), in which the scatter in $\delta_{\rm L}$ at fixed $\delta_{\rm E}$ is proportional to $S \equiv \sqrt{1 - r_{\rm LN}^2}$. This is a more meaningful quantity than the

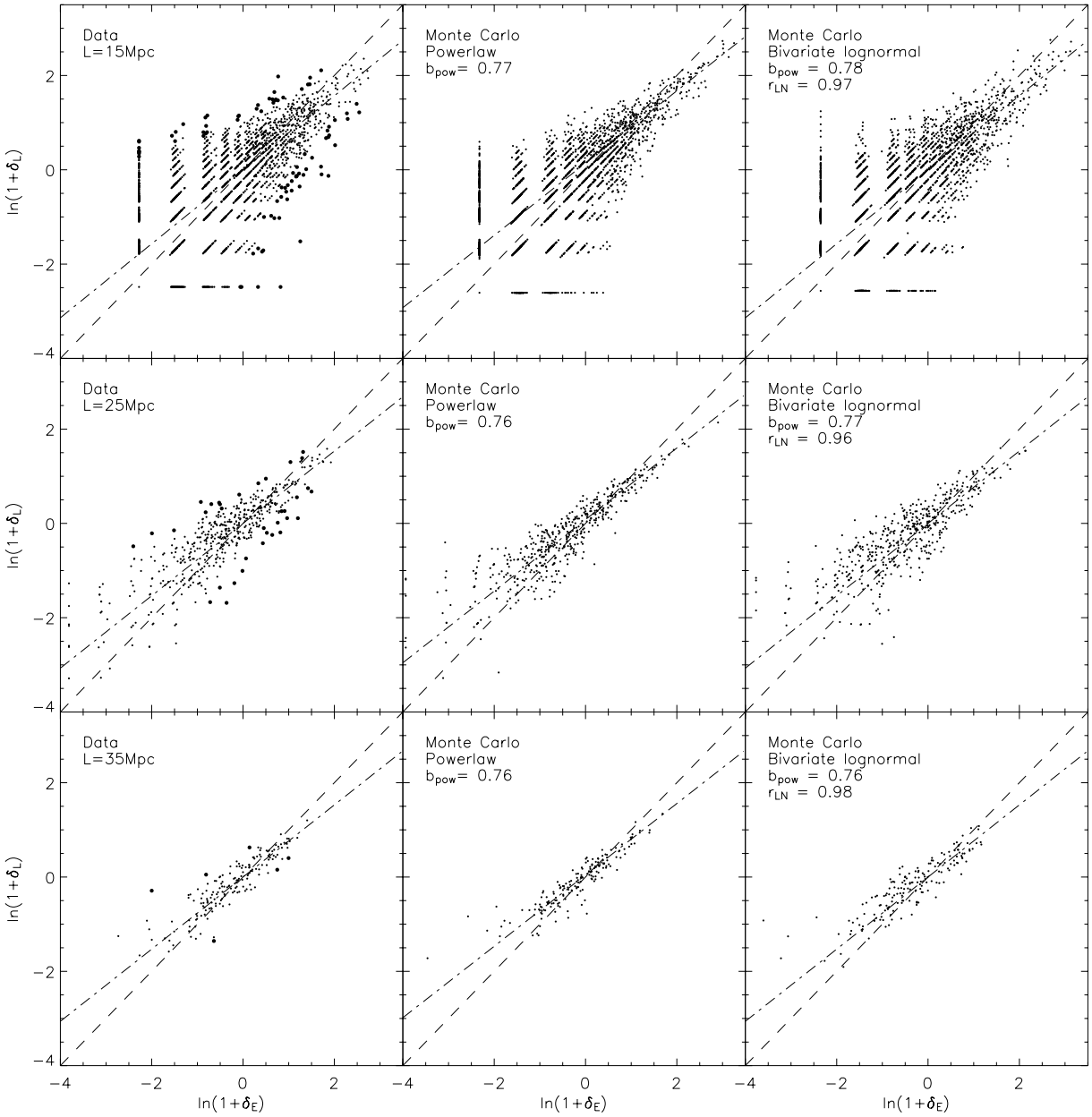


**Figure 9.** The bivariate counts-in-cells distributions for 15- (top), 25- and 35-Mpc length cells. The left-hand panel shows the data with early and late galaxies classified by colour. Larger points indicate the cells identified as outliers from $r_{\rm LN} = 1$ (see Section 6.3). The central column shows a Monte Carlo simulation of the best-fitting power-law model and the right-hand column shows the best-fitting bivariate lognormal model. The dashed line indicates a mean biasing of $b = 1.0$ and the dot-dash line shows the best-fitting mean bias.

correlation coefficient, but it lacks a standard name. In this context, the obvious term for $S$ would be 'stochasticity', but this is already taken and we resist the temptation to expand the terminology further. The stretched nature of this measure of correlation is quite extreme (as noted independently by Seljak & Warren 2004): $S = 0.5$ corresponds to $r_{\rm LN} = 0.87$. Therefore, $r_{\rm LN} = 0.87$ is effectively half-way to no correlation at all. This is why even a correlation as high as $r_{\rm LN} = 0.97$ is noticeably imperfect in terms of density–density plots.

### *6.2.3 Scale dependence*

We now look at how our results depend on scale. This is interesting because it may potentially distinguish whether the efficiency of galaxy formation in a particular region of space is affected by local or non-local factors. Examples of local factors could be density, geometry, or velocity dispersion of the dark matter. Non-local factors could involve, for example, effects of ionizing radiation from the first stars or QSOs on galaxy formation efficiency, causing coherent variation over larger scales than possible from local factors.

Fig. 9 shows the bivariate distributions for 15-, 25- and 35-Mpc cells with galaxies split by colour. The likelihood ratio tests (Table 3) show that the bivariate lognormal bias model provides a significantly better fit to our data than both deterministic models on all scales for colour selection and all but the largest scales when classifying by $\eta$.

The non-linearity and stochasticity as a function of scale are plotted in Fig. 10, with errors derived by propagation from those shown

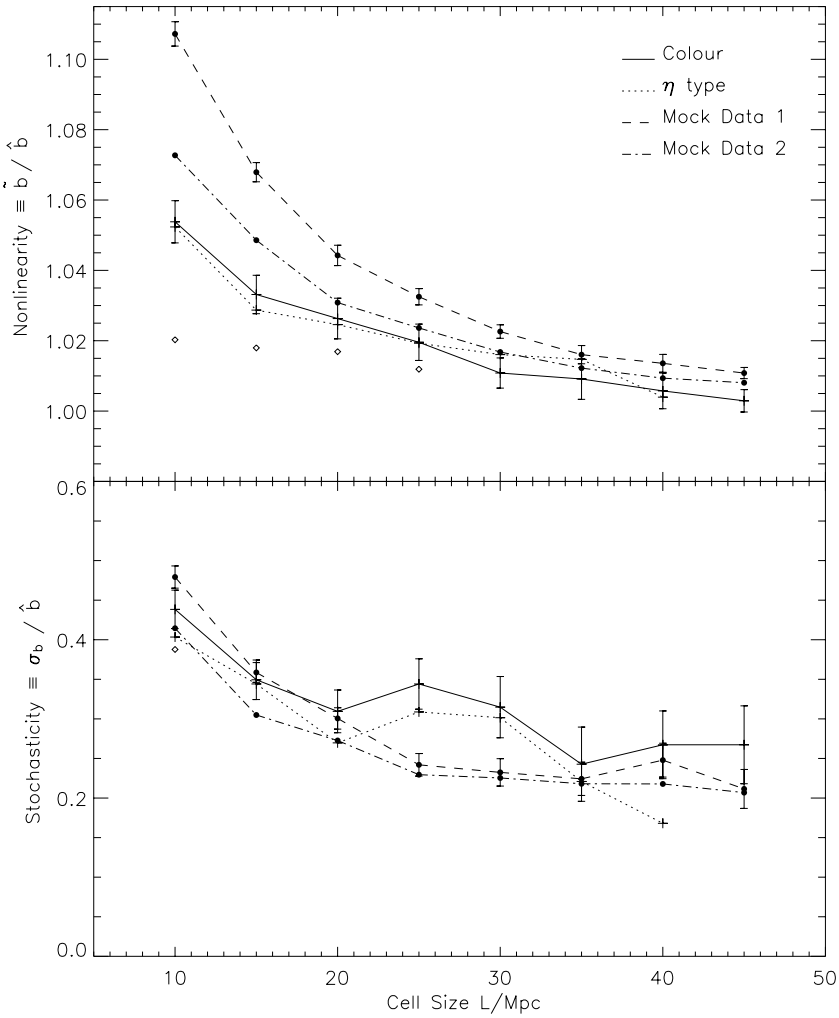


**Figure 10.** The scale dependence of non-linearity and stochasticity in the 2dFGRS. The solid line shows results for galaxies classified by colour and dotted line for galaxies classified by $\eta$ type. The circles show the two semi-analytic data sets. Mock 1 is described in the text as the 'superwind' model, and mock 2 as the 'low-baryon' model. The open diamonds indicate values measured for the colour data set using direct variance estimates, where they differ by more than $1\sigma$ from the results derived from model fitting (see Section 6.5). For clarity, errors are omitted for the $\eta$ and second mock data sets.

in Table 4. The commonly quoted parameters $b_{\rm var}$ and $r_{\rm lin}$, which combine both non-linearity and stochasticity, are plotted in Fig. 11 to ease reference with results in the literature. On small scales ($\leqslant$20 Mpc) the average biasing statistics suffer systematic errors from overestimates of the variance by the Poisson-sampled lognormal model fit, as discussed in detail in Section 6.5. To indicate the magnitude of these effects, open diamonds show results for the colour data set using direct variance estimates, where the difference is greater than $1\sigma$. It can be seen that although both non-linearity and $b_{\rm var}$ show noticeable change with scale, this can be mostly explained by the poor fit of the model. There is little effect on stochasticity, and both mocks and data are affected in the same way, making comparison practical. The non-linearity reaches $<$1 per cent by around 35 Mpc with results for colour and $\eta$ classification barely distinguishable. However, a little care is needed in interpreting this result: negligible 'non-linearity' does not mean that linear bias is a good fit. As much as anything, this is a statement that the amplitude of fluctuations declines for large $L$, so most cells have $|\delta| \ll 1$.

The stochasticity also declines, although on large scales the errors prevent distinction between a flat or declining function with scale. There is a tendency for the stochasticity of the $\eta$ data sets to lie a little below that of the colour data sets, but this is not significant within the errors. The dashed and dash-dot lines show the results for two semi-analytic mock universes which will be discussed in detail in Section 7.2. We can immediately note the encouraging general agreement: stochasticity is clearly expected at approximately the detected level.

### *6.2.4 Division by luminosity and redshift*

By splitting galaxies by their luminosity we can investigate whether the effects that we find in the previous sections could be due to the luminosity difference of the galaxy types. By dividing galaxies at $M - 5\log_{10}(h) = -19.5$, we form two similar sized groups with Class 1 being more luminous than Class 2. The models are fitted as before by replacing E (L) with Class 1 (2). In contrast to the outcome when galaxies are divided by type, the likelihood ratios between the best-fitting power-law and bivariate lognormal models are small on all scales, ranging from 0 to 3. We find $r_{\rm LN}$ to be roughly constant with a value of $\sim$0.99 for the best-fitting bivariate lognormal models. If stochasticity is caused by some variable other than the local density during galaxy formation, then perhaps luminosity

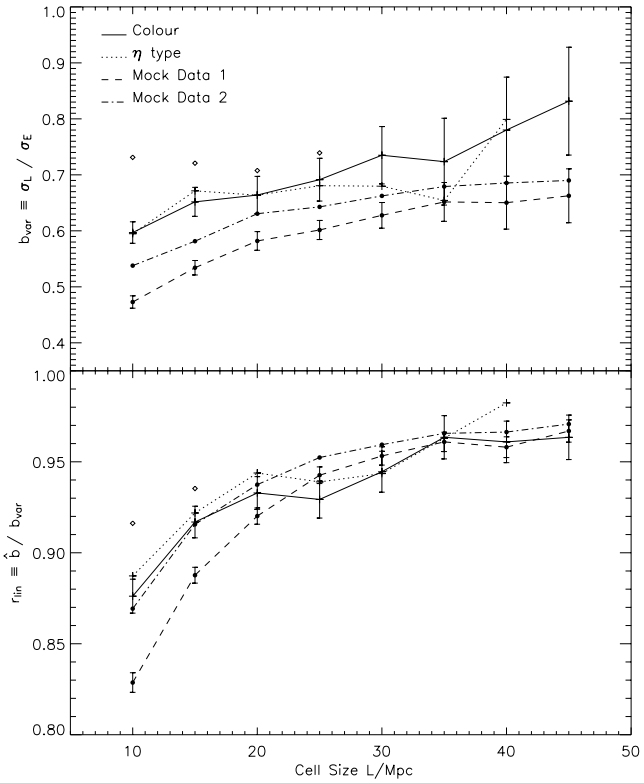


**Figure 11.** The scale dependence of the ratio of variances, $b_{\rm var}$, and the linear correlation coefficient, $r_{\rm lin}$. Symbols as in Fig. 10.

**Table 5.** Colour and $\eta$ samples with cell length $L = 20$ Mpc are each split into two redshift groups at $z = 0.09$. This table shows results for the bivariate lognormal model fit to each subsample.

| | | $\omega_{\rm E}$ | $\Delta(\omega_{\rm E})$ | $\omega_{\rm L}$ | $\Delta(\omega_{\rm L})$ | $r_{\rm LN}$ | $\Delta(r_{\rm LN})$ |
|---|---|---|---|---|---|---|---|
| Colour | Low $z$ | 1.11 | 0.03 | 0.85 | 0.03 | 0.956 | 0.010 |
| | High $z$ | 1.11 | 0.03 | 0.85 | 0.03 | 0.974 | 0.006 |
| $\eta$ | Low $z$ | 1.10 | 0.03 | 0.85 | 0.03 | 0.985 | 0.007 |
| | High $z$ | 1.07 | 0.03 | 0.82 | 0.03 | 0.976 | 0.007 |

is less dependent on this variable than galaxy type and colour. Other explanations could be that our volume-limited sample is too shallow to find the expected bimodality in luminosity, and the position of our boundary between bright and faint galaxies is arbitrary.

It is also of interest to see if the results are independent of redshift. We divide the survey at $z = 0.09$ and fit the models to both high- and low-redshift galaxies using a cell length of $L = 20$ Mpc. Table 5 shows the best-fitting bivariate lognormal parameters for galaxies split by colour and $\eta$ for both redshift groups. As a result of the fibre apertures of the 2dF instrument we may expect some redshift dependence for galaxies classified by $\eta$ (see Section 3.2.1), yet precise predictions are difficult. We certainly see no difference within the errors between these two redshift groups, and the difference for colour classification cannot be attributed to such effects. It is possible that the changing errors on the colour at high redshift contribute to the decrease in stochasticity, although evolution cannot be ruled out. There is certainly room for further investigation with forthcoming larger redshift surveys.

#### *6.2.5 Comparison with other 2dFGRS results*

This work has been carried out in conjunction with that of Conway et al. (2004) who investigate the variance and deviation from linear bias in the 2dFGRS NGP and SGP regions using flux-limited samples, including a counts-in-cells analysis. They find similar discrepancies between the Poisson-sampled lognormal model and the data, investigating the causes and magnitude of the problem in detail. After accounting for this effect in both analyses the results agree within the $1\sigma$ errors where comparable: Conway et al. find $1/b_{\rm var} = 1.25 \pm 0.05$, and non-linearity ($\tilde{b}/\hat{b}$) of a few per cent on the smallest scales measured. Our results for $b_{\rm pow}$ agree, but are consistently higher for $b_{\rm lin}$. This is due to different fitting procedures; Conway et al. give greater weight to overdense regions.

Madgwick et al. (2003a) measure the square root of the ratio of the correlation functions of early- and late-type galaxies to be around 1.2 on their largest scales of $8 < r < 20\,h^{-1}$ Mpc. Their bias parameter corresponds to $1/\sqrt{\hat{b}}$ in the notation of Section 2.1 (DL99). This gives a value for $\hat{b}$ a little higher than our results of Table 4, but which is entirely consistent when lognormal variance estimates are replaced by direct measures as in Section 6.2.3

### 6.3 Origin of the stochasticity signal

Before the detection of stochastic bias is accepted, and we proceed to confront the result with theoretical models, a degree of scepticism is in order. We have seen that some regions of space have a number ratio of early- and late-type galaxies that differs from the typical value by too much to be consistent with Poisson scatter. Such an outcome seems potentially vulnerable to systematics in the analysis as any source of error in classifying galaxies could introduce extra scatter, spuriously generating the impression of stochasticity. However, it is not clear which way this effect would go. Suppose the survey finds galaxies with perfect efficiency, but then assigns them a random class. Any true initial stochasticity is erased by the classification 'errors' and we measure $r_{\rm LN} = 1$. In order to generate apparent stochasticity where none is present we would need something more subtle. Possibilities could include a perfect efficiency in detecting early-type galaxies, but a fluctuating efficiency in finding late types; a spatially varying boundary between early and late types; or large variations in the survey selection function on scales smaller than the cell length. To assess the possible contribution of this latter effect to our measured stochasticity, we applied small-scale incompleteness masks to semi-analytic data sets (see Section 7.2). The large-scale stochasticity of these models was affected by less than $1\sigma$.

Whether or not a spurious generation of stochasticity seems plausible, it is worth looking more closely at the data to see how the signal arises. In order to do this, we focus on the outliers from the relation $\ln(1 + \delta_{\rm L}) \propto \ln(1 + \delta_{\rm E})$, but a careful definition of an outlier is required. We want to ask how much the numbers ($N_{\rm E}$, $N_{\rm L}$) differ from their expectation values when clustering is included, but the latter are unknown. Therefore, we take the best-fitting power-law model with $r_{\rm LN} = 1$ and integrate over the distribution of densities to calculate the probability for obtaining this outcome, $P(N_{\rm E}, N_{\rm L})$, accounting for Poisson noise. The most outlying points are those with the lowest values of $P$, and we remove these in succession until the remaining cells are consistent with an $r_{\rm LN} = 1$ model. The numbers of outliers in this sense are listed in Table 3, and Fig. 9 shows their positions on density–density plots.

Having identified the cells that provide the evidence for stochasticity, we can examine their properties in more detail. Fig. 12 shows the spatial distribution of the outlying cells within the 2dFGRS for

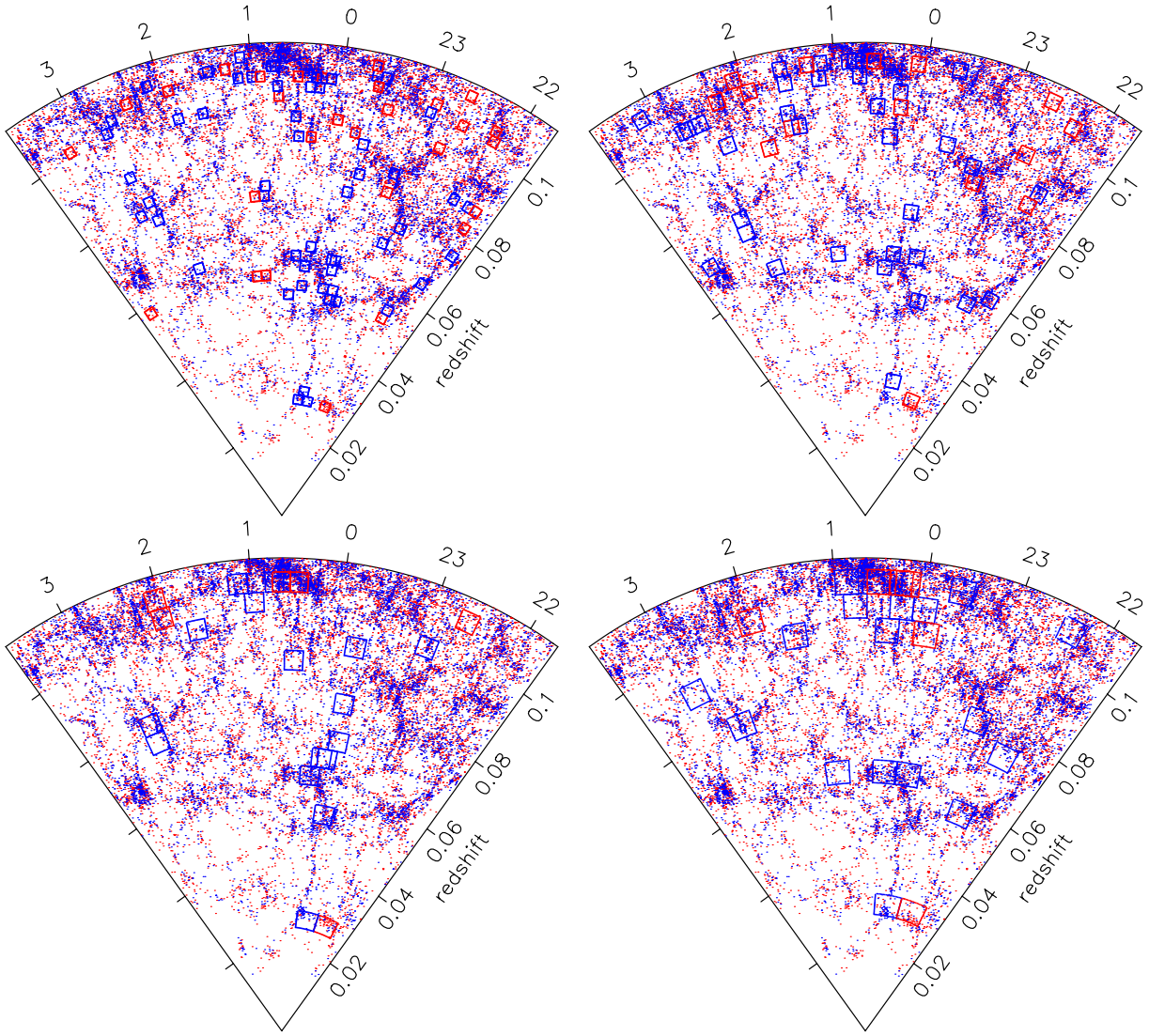


**Figure 12.** Wedge plots of the 2dFGRS volume-limited survey SGP region as for Fig. 3. Both early- and late-type galaxies are shown. Overplotted are cells identified as causing the stochasticity signal, from thetop left: 10-, 15-, 20-, 25-Mpc cells.

a range of cell sizes, from which it can be seen that they are often associated with overdense regions. This should not be taken as indicating that stochasticity is confined to such regions: given that the degree of stochasticity is small, the cells that contain the most galaxies will provide the best signal-to-noise ratio for the effect. The colour distribution of galaxies in the outlying cells is shown in Fig. 13, compared with the distribution of 'normal' cells. To allow for non-linearity in the density–density relation we consider for the comparison distribution only those cells with similar values of $\delta_{\rm E}$. The distributions cover sensible ranges of colours, and the peaks corresponding to early and late types appear to be in the correct places. What causes these cells to be outliers is that the ratio of the two populations differs greatly from what is typical, and it is hard to see how this result can be in error. The completeness values in these cells are typically 0.8, and yet we see variations in the early:late ratio by more than a factor of 2. Moreover, similar variations are seen whether we classify using colour or spectral type. We therefore conclude that these variations are a real property of the galaxy distribution.

### 6.4 Consistency checks

We repeat the analysis for cells with galaxies classified randomly, recovering a best-fitting bivariate lognormal model with $r_{\rm LN} = 1$ exactly. By fitting the bivariate lognormal model to Monte Carlo simulated power-law mocks (Section 5.3), we can check for any bias inherent in our fitting procedures. The best-fitting models have mean $r_{\rm LN} \gtrsim 0.998$, which is not significantly different from the $r_{\rm LN} = 1$ of power-law deterministic bias.

### 6.5 Direct variance estimates

It is possible to determine the variance $\sigma^2(L)$ directly without assuming the lognormal model. Optimal power spectrum estimates perhaps provide the most accurate determinations of variance (Tegmark et al. 2004; Pen et al. 2003), however, for our purposes it suffices to use a simpler method presented by Efstathiou et al. (1990). Their estimator calculates $\Delta N = N - \langle N \rangle$ for each cell and subtracts the Poisson variance from $(\Delta N)^2$ to form an estimate of $\sigma^2$ for each

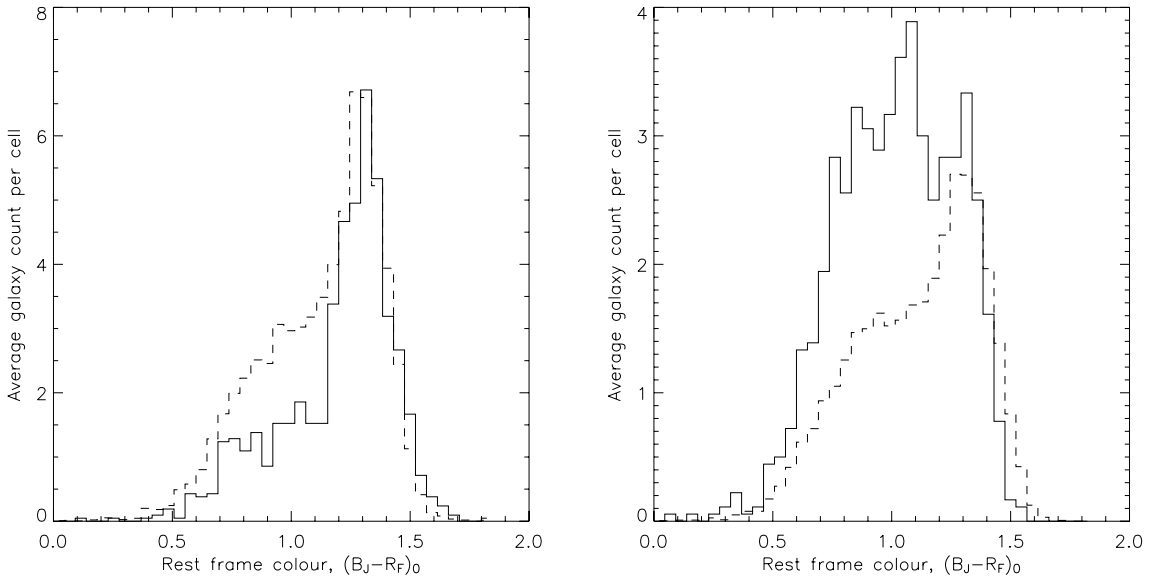


**Figure 13.** The colour distribution of cells identified as causing the stochasticity signal ($L = 20$ Mpc, filled line). The left (right) plot shows all those cells with an excess of red (blue) galaxies. The comparison plot (dashed line) is calculated from those cells with similar $\delta_E$ to the outlier cells, in order to account for non-linearities.

**Table 6.** Different variance estimates and errors for early and late data sets defined by colour. (*a*) From our bivariate lognormal model fit with errors derived from a multidimensional Gaussian fit to the likelihood surface; (*b*) Efstathiou et al. (1990) direct variance estimator and errors; (*c*) direct variance estimator after using Monte Carlo simulations of lognormal fields to correct for bias due to non-Gaussianity, with rms errors from the simulations.

| Cell size | $\sigma_E{}^a$ | $\Delta(\sigma_E)^a$ | $\sigma_E{}^b$ | $\Delta(\sigma_E)^b$ | $\sigma_E{}^c$ | $\Delta(\sigma_E)^c$ | $\sigma_L{}^a$ | $\Delta(\sigma_L)^a$ | $\sigma_L{}^b$ | $\Delta(\sigma_L)^b$ | $\sigma_L{}^c$ | $\Delta(\sigma_L)^c$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 3.014 | 0.074 | 1.806 | 0.016 | 1.814 | 0.142 | 1.799 | 0.037 | 1.321 | 0.013 | 1.323 | 0.059 |
| 15 | 1.987 | 0.060 | 1.462 | 0.021 | 1.472 | 0.129 | 1.295 | 0.033 | 1.054 | 0.016 | 1.056 | 0.056 |
| 20 | 1.542 | 0.057 | 1.242 | 0.028 | 1.253 | 0.126 | 1.024 | 0.035 | 0.879 | 0.021 | 0.882 | 0.055 |
| 25 | 1.214 | 0.048 | 1.008 | 0.034 | 1.019 | 0.107 | 0.839 | 0.032 | 0.746 | 0.026 | 0.750 | 0.054 |
| 30 | 0.922 | 0.045 | 0.877 | 0.042 | 0.890 | 0.102 | 0.677 | 0.033 | 0.659 | 0.032 | 0.663 | 0.061 |
| 35 | 0.807 | 0.064 | 0.728 | 0.041 | 0.736 | 0.087 | 0.584 | 0.042 | 0.547 | 0.032 | 0.551 | 0.050 |
| 40 | 0.764 | 0.067 | 0.716 | 0.049 | 0.727 | 0.095 | 0.596 | 0.049 | 0.557 | 0.038 | 0.563 | 0.061 |
| 45 | 0.657 | 0.054 | 0.547 | 0.047 | 0.557 | 0.071 | 0.546 | 0.044 | 0.467 | 0.041 | 0.473 | 0.058 |

cell. This is then averaged over all cells. The estimator only applies in the case of a uniform survey, where $\langle N \rangle$ is the same for all cells. For the general case of an incomplete survey, Efstathiou et al. derive a slightly different estimator, assuming a Gaussian density field. In fact, this is a poor assumption even on the largest scales considered here. We use Monte Carlo realizations of lognormal fields to show that their estimator for $\sigma$ is biased low by around 1–2 per cent, and has an uncertainty often several times larger than that expected for the Gaussian model. Table 6 gives the direct variance estimates for our data, with errors from both Efstathiou et al. (1990) and Monte Carlo simulations.

Even accounting for the bias, these direct variance estimates remain generally 10–20 per cent lower than those estimated by fitting a Poisson-sampled lognormal curve. For early-type galaxies in $L = 10$-Mpc cells the discrepancy is nearly 40 per cent. Imposing different weighting schemes during fitting can lower the lognormal variance to meet the direct variance results (Conway et al. 2004), but these have no significant effect on our measurement of stochasticity.

This failure of the lognormal model to recover the true variance of the data may be due to the assumption of the *Poisson clustering hypothesis* which we know to be incorrect in detail. On the smallest scales our cells are largely shot-noise-dominated, and it is on these scales that the discrepancy is greatest (see also Section 6.1.1). It remains important to emphasize that the variance estimates given throughout this paper are model dependent, and not to be taken as the true variance of the galaxies in the survey, which can be estimated more accurately through model-independent methods.

An unfortunate side effect of this difficulty in obtaining accurate estimates for variance, is that the average biasing statistics of Section 2.1 are dependent on $\sigma$ (see also Section 6.2.3). Tests show this to have little effect on stochasticity ($\sigma_b/\hat{b}$), but on scales $\leqslant$20 Mpc the non-linearity is overestimated. By replacing our measured variance with results obtained from bias-corrected direct estimation we find the non-linearity to decrease to around 2 per cent for $L = 10$ Mpc, decreasing with scale gradually to match our measured values by $L = 30$ Mpc. The stochasticity is decreased by approximately $2\sigma$ at $L = 10$ Mpc to around 0.39, but the effect is insignificant on all other scales. These results may be compared with the measurements of variance and deterministic bias in the 2dFGRS using flux-limited samples over a slightly larger volume (Conway et al. 2004).

This is a suitable point to discuss a subtlety of cell counts that we have neglected so far. Variation in the survey mask is represented by $\langle N \rangle$ varying between cells. We have treated this as a simple variation in sampling efficiency that is uniform over the cell. However, this cannot be precisely correct: where sampling of a cell

is low because it encounters one of the larger drills in the input catalogue, it would be more correct to assume a completely sampled cell of smaller volume. We have explored this alternative extreme by assuming that $\sigma \propto \langle N \rangle^{-0.3}$, as expected for a $\xi(r) \propto r^{-1.8}$ spectrum. As $\langle N \rangle \propto$ completeness, and the typical cell completeness is approximately 0.8, the measured values of $\sigma$ are increased by approximately 7 per cent, approximately a $1\sigma$ shift. This has no effect on our detection of stochasticity, and because the 'lost volume' assumption will not apply in all cases, we neglect the issue. Note that estimates of cell variances derived from integration over correlation functions or power spectra would be completely independent of this issue.

## 7 COMPARISON WITH SIMULATIONS

In order to interpret our measurements of stochastic bias, we need to make a comparison with theory. In practice, this means considering the results of numerical simulations that are sufficiently detailed to predict the spatial distributions of the different classes of galaxy. There are currently two main methods of simulating the large-scale structure of the visible Universe: semi-analytic or hydrodynamic. We consider each in turn.

### 7.1 Previous work

Somerville et al. (2001) used semi-analytic models to measure the relative bias between early- and late-type galaxies (as defined by bulge to total luminosity) and red and blue galaxies on scales of $r = 8\,h^{-1}$ Mpc. They set a limiting magnitude of $M_B - 5 \log h \leqslant -18.4$, and split galaxies by colour at $B - R = 0.8$, making their samples reasonably comparable to our data for cells of $L = 20$ Mpc. Our value of $b_{\rm var} = 0.66$ (Table 4) falls in between their values of 0.77 for late/early types and 0.55 for blue/red galaxies. They find $r_{\rm lin} = 0.87$ for both subgroups, slightly lower than our values for both colour and spectral type. Unfortunately the results are not split into stochasticity and non-linearity, making it difficult to make further comparisons. It is, however, interesting that we find a lower amplitude of relative bias between the two colour groups than is seen in these models.

The hydrodynamic simulations of Yoshikawa et al. (2001) classify galaxies by their formation redshift, and are smoothed with top-hat spheres of radius $8\,h^{-1}$ Mpc. By using this classification scheme, hydrodynamic models approximate early-type galaxies as those that form at high redshifts via initial starbursts, whereas late-type galaxies have a lower formation redshift and undergo slower star formation. They find that old galaxies are positively biased with respect to matter with a linear correlation coefficient of less than 1, whereas young galaxies are slightly antibiased with a correlation coefficient closer to 1. They measure the relative bias between galaxy types by $b_\xi^{\rm rel} \equiv (\xi_{\rm young}/\xi_{\rm old})^{1/2}$, where $\xi_{\rm young}(\xi_{\rm old})$ is the two-point correlation function of the young (old) galaxies. This is equivalent to $b_{\rm lin}$. They obtain values of between 0.5 and 0.66 for scales of $1 < r < 20\,h^{-1}$ Mpc, lower than our equivalent values for the linear biasing model with $L \leqslant 25$-Mpc cells (Table 3). Once again results for stochasticity and non-linearity are not quoted for the relative bias.

### 7.2 Preliminary mock comparison

None of this past work really allows a direct comparison with our results, so we generated two new theoretical 'data sets' from the results of large semi-analytic calculations carried out using the 'Cosmology machine' supercomputer at Durham. The background model is that deduced from the simplest *WMAP*+2dFGRS analysis of Spergel et al. (2003): flat, $\Omega_{\rm m} = 0.27$, $\Omega_{\rm b} = 0.045$, $h = 0.72$, $n = 0.97$, $\sigma_8 = 0.8$, applying the semi-analytic apparatus of Cole et al. (2000) to a simulation with $N = 500^3$ particles in a box of side $250\,h^{-1}$ Mpc. As shown by, for example, Benson et al. (2003), a problem faced by such modelling is a tendency to overproduce massive galaxies, as a result of excessive cooling arising from the higher baryon density now required by CMB+LSS. This problem is particularly severe for disc (late-type) galaxies. The first mock adopted the 'superwind' approach of Benson et al. (2003) in an attempt to alleviate this problem, but the cure is not total. The second mock attempted to reduce cooling by retaining the low baryon density of Cole et al. (2000). Although this conflicts with CMB data, it provides a useful means of comparison. For this application, we took an empirical approach in which a monotonic shift in luminosity was applied to force the models to have the observed luminosity function as in Madgwick et al. (2002). The model colour distribution was bimodal to a realistic degree, so this shift was applied separately to generate model distributions of early- and late-type galaxies in which the global luminosity functions were correct. The resultant mock cell counts were analysed identically to the real data.

In some respects, these simulations match the real data very well. For the low-baryon model, the amplitude of the cell variances for early-type galaxies agree to within 3 per cent on small scales and 10 per cent on large scales. The superwind model variances agree to within 10 per cent on all scales. The relative bias of the low-baryon model agrees to within 10 and 15 per cent with observation, and the superwind model to within 10 and 20 per cent. Significant stochasticity and non-linearity is also required, which can be measured accurately as a function of scale because we are able to use more mock cells than are available in the real data. The mocks are affected in a similar manner to the data by the discrepancy between direct estimates of variance and those from lognormal model fits. On small scales this significantly increases our estimated non-linearity; as the effect is equivalent between mocks and data, however, a direct comparison between them remains instructive. Fig. 10 shows the resulting stochasticity and non-linearity as a function of scale, compared with that of the 2dFGRS data. The impression is that the mock results show a greater non-linearity than the real data on small scales, while stochasticity is well matched within the errors.

Given the known imperfect nature of the semi-analytic simulations (e.g. the failure to match luminosity functions exactly), the correct attitude is probably to be encouraged by the degree of agreement with the data. It is certainly plausible that the existing calculations contain all the relevant physical contributions to bias, but perhaps not yet in quite the right proportions. As is usual with such numerical comparisons, this raises the question of whether the issue of stochasticity can be understood in a more direct fashion. In the end, the effects we are seeing must be reducible to the way in which the early:late ratio varies between and within virialized systems of different mass, so that in effect we are dealing with a more general version of the morphological segregation that is familiar from the study of rich clusters (Narayanan, Berlind & Weinberg 2000). We intend to pursue this in more detail elsewhere, using the catalogue of galaxy groups derived from the 2dFGRS by Eke et al. (2004).

## 8 SUMMARY AND CONCLUSIONS

We have presented fits of three relative biasing schemes to joint counts-in-cells distributions of 2dFGRS galaxies, separated by both colour and spectral type $\eta$. Each scheme is convolved with a Poisson distribution to account for statistical 'shot noise'. Our first two models present two alternative types of deterministic biasing:

linear and power-law bias. Linear bias is an important concept in cosmology and many results are linked to it, but it is not physically plausible as it allows negative densities. Power-law bias presents a simple cure for this problem, but still has little physical motivation. With the advent of large semi-analytic and hydrodynamic simulations, interest has grown in 'stochastic' bias models. Bias could be determined by parameters other than the local overdensity of the dark matter, and considerable scatter could occur in the relation. Galaxy distributions have previously been measured to be well approximated as lognormal, therefore a bivariate lognormal distribution seems a natural model for relative bias between galaxy types. This model incorporates stochasticity and non-linearity in a well-defined manner, which is mathematically simple and consistent with observation.

To account for the discrete nature of galaxies, the *Poisson sampling hypothesis* is assumed, and all models are convolved with a Poisson distribution. On small scales where our cell counts become shot-noise-dominated, we find this hypothesis to fail, causing overestimates of variance compared with direct estimation methods. The main symptom of the discrepancy is a number of completely empty cells that exceeds the Poisson-sampled lognormal prediction. This is found not to affect our results for stochasticity, and the same effect is seen in the simulations, but it emphasizes the need for a greater understanding of Poisson statistics in relation to galaxy clustering.

We have detected a significant deviation from $r_{\rm LN} = 1$ in the 2dFGRS and confirmed this detection of stochasticity through likelihood ratio tests, Kolmogorov–Smirnoff probability testing and Monte Carlo simulations. We have measured stochasticity at a level of $\sigma_{\rm b}/\hat{b} = 0.44 \pm 0.02$ or $r_{\rm LN} = 0.958 \pm 0.004$ on the smallest scales (10 Mpc), declining with increasing cell size. The non-linearity of the biasing relation is less than 5 per cent on all scales. The small measured values of stochasticity and non-linearity support the use of galaxy redshift surveys for studies of the large-scale distribution of matter in the Universe, and the measurement of cosmological parameters. However, as precision in cosmology increases and new techniques are developed, the effects of stochastic bias on parameter estimation should be understood. For example, studies of cosmology through weak gravitational lensing requires knowledge of non-linear and stochastic bias (Seljak & Warren 2004). Our results for $r_{\rm lin}$ on 10-Mpc scales are consistent within the (large) errors with galaxy–mass correlations measured by weak-lensing surveys (Hoekstra et al. 2002) on the largest scales probed.

A comparison with semi-analytic simulations shows a similar variation of non-linearity and stochasticity with scale. The amplitude of stochasticity appears to be a little lower than in the true data, particularly on large scales, and the non-linearity is slightly greater on small scales. Nevertheless, given the known imperfections of the current generation of semi-analytic calculations, the general agreement is certainly encouraging. We hope that this work will stimulate the investigation of more detailed biasing models. Through the linking of new simulations to observations, a more thorough understanding of the processes of galaxy formation and evolution should be within our reach.

## ACKNOWLEDGMENTS

The 2dF Galaxy Redshift Survey was made possible through the dedicated efforts of the staff of the Anglo-Australian Observatory, both in creating the 2dF instrument and in supporting it on the telescope. JAP and OL thank PPARC for their Senior Research Fellowships. Many thanks go to Paul Hewett, Yehuda Hoffman, Mike Irwin and Dan Mortlock for helpful discussions on galaxy evolution and statistics.

## REFERENCES

Baldry I. K., Glazebrook K., Brinkmann J., Ivezic Z., Lupton R. H., Nichol R. C., Szalay A. S., 2004, ApJ, 600, 681
Bardeen J. M., Bond J. R., Kaiser N., Szalay A. S., 1986, ApJ, 304, 15
Baugh C. M. et al. (the 2dFGRS Team), 2004, MNRAS, 351, L44
Benson A. J., Baugh C. M., Cole S., Frenk C. S., Lacey C. G., 2000, MNRAS, 316, 107
Benson A. J., Bower R. G., Frenk C. S., Lacey C. G., Baugh C. M., Cole S., 2003, ApJ, 599, 38
Bernardeau F., Kofman L., 1995, ApJ, 443, 479
Blanton M., 2000, ApJ, 544, 63
Blanton M., Cen R., Ostriker J. P., Strauss M. A., 1999, ApJ, 522, 590
Bower R. G., Coles P., Frenk C. S., White S. D. M., 1993, ApJ, 405, 403
Broadhurst T. J., Taylor A. N., Peacock J. A., 1995, ApJ, 438, 49
Cardelli J. A., Clayton G. C., Mathis J. S., 1989, ApJ, 345, 245
Cen R., Ostriker J. P., 1992, ApJ, 399, L113
Cen R., Ostriker J. P., 2000, ApJ, 538, 83
Cole S., Kaiser N., 1989, MNRAS, 237, 1127
Cole S., Lacey C. G., Baugh C. M., Frenk C. S., 2000, MNRAS, 319, 168
Coles P., 1993, MNRAS, 262, 1065
Coles P., Jones B., 1991, MNRAS, 248, 1
Colless M. et al. (the 2dFGRS Team), 2001, MNRAS, 328, 1039
Colless M. et al. (the 2dFGRS Team), 2003, preprint (astro-ph/0306581)
Conway E. et al. (the 2dFGRS Team), 2004, MNRAS, in press (astro-ph/0404276)
Cross N. J. G., Driver S. P., Liske J., Lemon D. J., Peacock J. A., Cole S., Norberg P., Sutherland W. J., 2004, MNRAS, 349, 576
Croton D. J. et al. (the 2dFGRS Team), 2004a, MNRAS, 352, 828
Croton D. J. et al. (the 2dFGRS Team), 2004b, MNRAS, 352, 1232
Davis M., Geller M. J., 1976, ApJ, 208, 13
Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, ApJ, 292, 371
Davis M. et al. (the DEEP2 Team), 2003, in Guhathakurta P., ed., Proc. SPIE, Vol. 4834, Discoveries and Research Prospects from 6- to 10-Meter-Class Telescopes II. Int. Soc. Opt. Eng., Bellingham, WA, p. 161
Dekel A., Lahav O., 1999, ApJ, 520, 24 (DL99)
Dekel A., Ostriker J. P., 1999, Formation of Structure in the Universe. Cambridge Univ. Press, Cambridge
Dressler A., 1980, ApJ, 236, 351
Efstathiou G., Kaiser N., Saunders W., Lawrence A., Rowan-Robinson M., Ellis R. S., Frenk C. S., 1990, MNRAS, 247, 10
Eke V. R. et al. (the 2dFGRS Team), 2004, MNRAS, 348, 866
Elgarøy Ø., Lahav O., 2003, J. Cosmol. Astro-Particle Phys., 4, 4
Fan Z., 2003, ApJ, 594, 33
Fasano G., Franceschini A., 1987, MNRAS, 225, 155
Fischer P. et al. (the SDSS Collaboration), 2000, AJ, 120, 1198
Folkes S. et al. (the 2dFGRS Team), 1999, MNRAS, 308, 459
Fry J. N., Gaztanaga E., 1993, ApJ, 413, 447
Gunn J. E., Gott J. R. I., 1972, ApJ, 176, 1
Hambly N. C. et al., 2001, MNRAS, 326, 1279
Hamilton A. J. S., 1985, ApJ, 292, 35
Hawkins E. et al. (the 2dFGRS Team), 2003, MNRAS, 346, 78
Helly J. C., Cole S., Frenk C. S., Baugh C. M., Benson A., Lacey C., Pearce F. R., 2003, MNRAS, 338, 913
Hermit S., Santiago B. X., Lahav O., Strauss M. A., Davis M., Dressler A., Huchra J. P., 1996, MNRAS, 283, 709
Hoekstra H., van Waerbeke L., Gladders M. D., Mellier Y., Yee H. K. C., 2002, ApJ, 577, 604
Hubble E., Humason M. L., 1931, ApJ, 74, 43
Kaiser N., 1987, MNRAS, 227, 1
Kauffmann G., Nusser A., Steinmetz M., 1997, MNRAS, 286, 795
Kayo I., Taruya A., Suto Y., 2001, ApJ, 561, 22
Kofman L., Bertschinger E., Gelb J. M., Nusser A., Dekel A., 1994, ApJ, 420, 44

Lahav O., Saslaw W. C., 1992, ApJ, 396, 430
Lahav O., Nemiroff R. J., Piran T., 1990, ApJ, 350, 119
Lahav O. et al. (the 2dFGRS Team), 2003, MNRAS, 333, 961
Liddle A. R., 2004, MNRAS, 351, L49
Lupton R., 1993, Statistics in Theory and Practice. Princeton University Press, Princeton
Maddox S. J., Efstathiou G., Sutherland W. J., Loveday J., 1990, MNRAS, 243, 692
Madgwick D. S., 2003, MNRAS, 338, 197
Madgwick D. S. et al. (the 2dFGRS Team), 2002, MNRAS, 333, 133
Madgwick D. S. et al. (the 2dFGRS Team), 2003a, MNRAS, 344, 847
Madgwick D. S., Somerville R., Lahav O., Ellis R., 2003b, MNRAS, 343, 871
Matsubara T., 1999, ApJ, 525, 543
Narayanan V. K., Berlind A. A., Weinberg D. H., 2000, ApJ, 528, 1
Norberg P. et al. (the 2dFGRS Team), 2001, MNRAS, 328, 64
Norberg P. et al. (the 2dFGRS Team), 2002, MNRAS, 332, 827
Peacock J. A., 1983, MNRAS, 202, 615
Peacock J. A., 2003, in Holt S. H., Reynolds C. S., eds, AIP Conf. Proc., Vol. 666, The Emergence of Cosmic Structure. Am. Inst. Phys., New York, p. 275
Peacock J. A. et al. (the 2dFGRS Team), 2001, Nat, 410, 169
Peebles P. J. E., 1980, The Large-scale Structure of the Universe. Princeton University Press, Princeton, p. 435
Pen U., 1998, ApJ, 504, 601
Pen U., Lu T., van Waerbeke L., Mellier Y., 2003, MNRAS, 346, 994
Percival W. J. et al. (the 2dFGRS Team), 2001, MNRAS, 327, 1297
Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P., 1992, Numerical Recipes in C. The Art of Scientific Computing, 2nd edn. Cambridge University Press, Cambridge
Scherrer R. J., Weinberg D. H., 1998, ApJ, 504, 607
Schlegel D. J., Finkbeiner D. P., Davis M., 1998, ApJ, 500, 525
Seljak U., Warren S., 2004, MNRAS, in press (astro-ph/0403698)
Sheth R. K., Mo H. J., Saslaw W. C., 1994, ApJ, 427, 562
Somerville R. S., Primack J. R., 1999, MNRAS, 310, 1087
Somerville R. S., Lemson G., Sigad Y., Dekel A., Kauffmann G., White S. D. M., 2001, MNRAS, 320, 289
Spergel D. N. et al. (the WMAP Team), 2003, ApJS, 148, 175
Spitzer L. J., Baade W., 1951, ApJ, 113, 413
Strauss M. A. et al. (the SDSS Collaboration), 2002, AJ, 124, 1810
Tegmark M., Bromley B. C., 1999, ApJ, 518, L69
Tegmark M. et al. (the SDSS Collaboration), 2004, ApJ, 606, 702
Valageas P., Munshi D., 2004, MNRAS, in press (astro-ph/0403593)
Verde L. et al. (the 2dFGRS Team), 2002, MNRAS, 335, 432
Verde L. et al. (the WMAP Team), 2003, ApJS, 148, 195
Willmer C. N. A., da Costa L. N., Pellegrini P. S., 1998, AJ, 115, 869
Yoshida N., Stoehr F., Springel V., White S. D. M., 2002, MNRAS, 335, 762
Yoshikawa K., Taruya A., Jing Y. P., Suto Y., 2001, ApJ, 558, 520
Zehavi I. et al. (The SDSS Collaboration), 2002, ApJ, 571, 172
Zehavi I. et al. (the SDSS Collaboration), 2004, ApJ, 608, 16

## APPENDIX A: NON-LINEAR AND STOCHASTIC BIAS STATISTICS FOR THE BIVARIATE LOGNORMAL DISTRIBUTION

The general biasing relation between the overdensities $\delta_1$ and $\delta_2$ of two subgroups of galaxies, or two types of matter, is fully described by equation (7)

$$b(\delta_1)\delta_1 \equiv \langle \delta_2|\delta_1\rangle = \int f(\delta_2|\delta_1)\delta_2 \, \mathrm{d}\delta_2. \tag{A1}$$

The conditional probability distribution for the bivariate lognormal model is given by equation (33)

$$f(g_2|g_1) = \frac{\omega_1}{(2\pi|V|)^{1/2}} \exp\left[-\frac{(\tilde{g_2} - r_{\rm LN}\,\tilde{g_1})^2}{2\left(1-r_{\rm LN}^2\right)}\right], \tag{A2}$$

where $g_i = \ln(1+\delta_i) - \langle \ln(1+\delta_i)\rangle$, $\langle \ln(1+\delta_i)\rangle = -\omega_i^2/2$ and $\tilde{g_i} = g_i/\omega_i$. $\tilde{g_2}|\tilde{g_1}$ follows a univariate Gaussian distribution with mean $r_{\rm LN}\,\tilde{g_1}$ and variance $1-r_{\rm LN}^2$. The covariance matrix $V$ and correlation coefficient $r_{\rm LN}$ are both defined in log space, and are given explicitly in equations (31) and (32). The variance of the distribution in linear space $\sigma_i^2$ is related to the variance of the Gaussian field by

$$\sigma_i^2 \equiv \left\langle \delta_i^2 \right\rangle = \exp\left(\omega_i^2\right) - 1. \tag{A3}$$

On substituting equation (A2) into (A1) and integrating we find

$$b(\delta_1)\delta_1 = \exp\left[\omega_2\tilde{g_1} - \frac{(\omega_2 r_{\rm LN})^2}{2}\right] - 1. \tag{A4}$$

From this basic parameter we can calculate the mean biasing and its non-linearity (equations 8 and 9)

$$\hat{b} \equiv \frac{\left\langle b(\delta_1)\delta_1^2\right\rangle}{\sigma_1^2} = \frac{\exp(r_{\rm LN}\omega_2\omega_1) - 1}{\exp\left(\omega_1^2\right) - 1}, \tag{A5}$$

$$\tilde{b}^2 \equiv \frac{\left\langle b^2(\delta_1)\delta_1^2\right\rangle}{\sigma_1^2} = \frac{\exp\left(r_{\rm LN}^2\omega_2^2\right) - 1}{\exp\left(\omega_1^2\right) - 1}. \tag{A6}$$

We also know the ratio of variances, equation (15)

$$b_{\rm var}^2 \equiv \frac{\sigma_2^2}{\sigma_1^2} = \frac{\exp\left(\omega_2^2\right) - 1}{\exp\left(\omega_1^2\right) - 1}. \tag{A7}$$

Although it is possible to derive the scatter for this model from equation (12), it can be shown that (DL99)

$$b_{\rm var}^2 = \tilde{b}^2 + \sigma_{\rm b}^2. \tag{A8}$$

Using this fact we obtain

$$\sigma_{\rm b}^2 = \frac{\exp\left(\omega_2^2\right) - \exp\left(r_{\rm LN}^2\omega_2^2\right)}{\exp\left(\omega_1^2\right) - 1}. \tag{A9}$$

Whilst the bivariate lognormal model contains constant scatter dependent only on $r_{\rm LN}$ in the log frame, transformation to the linear frame causes the scatter to become dependent on the widths of the univariate distributions and vary with $\delta_1$.

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