A void perspective of the cosmic web
Platen, Erwin

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2009

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
5

The SDSS Density Field
Reconstruction, Higher Order
Methods and Nonlinearities

— E. Platen, M. A. Aragón Calvo, R. van de Weygaert, B.J.T. Jones & G. Vegter —

In this chapter we investigate the reconstruction of the density field from the Sloan Digital Sky Survey (SDSS) DR6 data releases. We compare the performance of three density field reconstruction methods. The first two are local geometry estimators, the linear Delaunay Triangulation Field Estimator (DTFE) and a higher order extension Natural Neighbour Field Estimator (NNFE). We introduce another method called the Natural Lognormal Kriging method, it is based on a global method which takes into account the spatial correlations within a naturally defined environment. We pay special attention to nonlinear sites and show how to take proper care of these features. The three methods are applied to a set of mock catalogues from the Millennium simulation. This provides us with direct insight in the behaviour of the errors. Using correlation diagrams we investigate the existence of possible biases in the methods. The tests were carried out at various scales and distances using both magnitude limited and volume limited samples. Using mock catalogues we provide limits on what can be recovered and what is lost in case of a SDSS density field reconstruction. Within a given distance the density field can be accurately reconstructed to a certain scale depending on the point intensity at that given distance. The samples may either be volume - or magnitude limited, in the latter case the errors are smaller, yet they are statistically nonuniform. In magnitude limited samples cosmic variance can introduce systematic deviation at local distances ($<100h^{-1}\text{Mpc}$), though these disappear at larger distances. All three methods were shown to be unbiased at all scales, producing very similar errors. The two higher order schemes produce visually much smoother fields. An analysis of the topology of the field in the underdense regime did not show any improvement over the linear DTFE method. We actually find that the simplest and computationally least demanding method DTFE is the most robust and has the best performance.

5.1 Introduction

The shape and statistics of the cosmological density field depend strongly on how structures in the Universe have been assembled. According to the standard lore of structure formation structures originated from small perturbations in the initial density field. This primordial field is described by a Gaussian Random Field. Under the force of gravity these fluctuations have grown and clustered to become the present day observed structures. At large scales the density field has been evolving (quasi)-linearly and still retains much of the primordial cosmological information. An abundance of probes have been developed that can retrieve this
Figure 5.1– A visualisation of the (DTFE) SDSS density field ($1h^{-1}\text{Mpc}$), the contour levels are divided roughly into the overdense and the underdense regime. Both the galaxies (blue dots) and the density represent a slice of $12h^{-1}\text{Mpc}$ thickness. Some of the most prominent features have been named, like the Boötes SuperVoid (Kirshner et al. 1981) and the large supervoid (BS SuperVoid) identified by Bahcall & Soneira (1982). Also the largest overdense structure, the (Coma) Great Wall, and the location of the Hercules supercluster within the Wall are indicated.

information, e.g. the galaxy clustering (Peebles 1980; Percival et al. 2001; Tegmark et al. 2004), the temperature and polarisation anisotropies in the CMB (Smoot et al. 1992; Spergel et al. 2003), identification of morphological structures (Zel’Dovich 1970; Bond et al. 1996) (voids, filaments, walls, superclusters, etc.) in the density field, weak lensing caused by the underlying inhomogeneous matter distribution (Mellier 1999; Massey et al. 2007; Hoekstra & Jain 2008) and many more. With this analysis we may recover the cosmological parameters or study the formation of large scale structure. The Megaparsec scale Universe represents an important cosmological probe.

On small scales the picture is more complicated. The full nonlinear growth has significantly altered and erased some of the essential cosmological information. To some extent fully collapsed haloes may be regarded as decoupled systems from the cosmological background. These systems flow along with the large scale displacement field until a new significant massive merger occurs. The mass and number density distribution of these peaks depend significantly on the ambient environment. Low density regions by definition contain fewer peaks, and they are also less massive than their overdense peers. Since peaks are the site where galaxies form, they are (biased) tracers of the underlying density field.

Our goal here is to reconstruct the underlying field as accurately as possible in order to probe the complex structure in the quasi-linear (and nonlinear) regime. This also includes the anisotropy of nonlinear structures. Often reconstruction is a first step in the process of retrieving the cosmological information (see the above mentioned examples). Therefore it is crucial to understand the details and errors in reconstructed density fields. We will provide a detailed comparison between the three reconstruction methods, and focus on the errors at a range of scales. From the smallest scale, at $1h^{-1}\text{Mpc}$, which probes the full nonlinear regime to the largest scale, $10h^{-1}\text{Mpc}$, that represents the transition from quasi-linear to the linear growth regime.

The typical data we consider are galaxy positions from modern (redshift/photometric)
Figure 5.2— The figure illustrates the interpolation process. Left column shows from top to bottom a surface plot (top) and a contour plot (middle) of a Gaussian test function that is sampled by 180 random points (bottom). The three columns to the right show the three methods that are reconstructed with DTFE (2nd column), NNFE (3rd) and the Kriging method (last column). In the middle row the red contours are the reconstructed belonging to the same isodensity levels as the original black contours. The bottom row shows the interpolation errors, as provided by the absolute difference between interpolated and original.

surveys like the two degree Field Galaxy Redshift Survey (2dFGRS), 6dFGRS, the Sloan Digital Sky Survey (SDSS) and other upcoming surveys. At the moment SDSS covers one of the largest and deepest contiguous regions of the nearby Universe. This makes SDSS an ideal survey to reconstruct the full 3-dimensional density field. Here we focus on a SDSS density reconstruction, but the overall behaviour of the methods is generic and does carry over to any other redshift survey. A first impression how such a density reconstruction may look like is provided in figure 5.1.

5.1.1 Reconstruction Methods

A multitude of approaches have been used to reconstruct the cosmological density field. For example, there are methods based on the physical relationship between peculiar velocity and the density. The POTENT method (Dekel et al. 1990) is a well-known example. A three dimensional reconstruction of the dark matter density field is also possible using weak gravitational
lensing (Kaiser 1992; Massey et al. 2007). The most common approach is the use of the observed spatial galaxy and cluster distribution to infer the underlying density field.

We pursue the reconstruction of a density field from a spatial point process, assuming that the local intensity of the points is a fair tracer of the density and that these values are samples of a continuous underlying density field. Within this framework, reconstruction consists of two essential steps. The first is the estimation of the local galaxy density, the second is the interpolation of the densities to obtain a continuous spatial field. Following straightforward grid based interpolation methods and more sophisticated adaptive filter techniques, there has been a substantial investment into developing more advanced techniques. Particular examples are the Wiener filter reconstruction (Zaroubi et al. 1995; Zaroubi 2002; Erdogdu et al. 2004, 2006; Kitaura & Enßlin 2008). An extensive overview of this technique is given in Kitaura & Enßlin (2008) and references therein. A recently introduced technique is the Wavelet reconstruction (Martínez et al. 2005) which takes into account the multiscale nature of the Cosmic Web.

A graphical representation of this outline can be seen in Figure 5.3 (see Kitaura & Enßlin 2008, for a slightly different outline). We will focus here mainly on the second step, the interpolation step. In order to compare the interpolation methods in an unbiased fashion, the same initial density estimates (first step) are used for each method.

Interpolation on scattered points aims to approximate a continuous function constrained by the available data points. This problem occurs in many branches of science. A wide variety of methods have been developed. Each having their own advantages and disadvantages. They can be roughly divided into two categories, global and local methods. Local methods have the benefit that they are fast and able to deal with large data-sets. Global methods produce smoother interpolated functions, yet are computationally more expensive. An introduction into the topic of spatial interpolation can be found in chapter 3.4 of Press (2007) and in Watson (1992). Lombardi (2002) provides an overview of the statistical properties of interpolation techniques, Franke (1982) and Amidror (2002) give detailed comparisons between various methods.
Table 5.1– Interpolation methods of scattered data and their characteristics.

<table>
<thead>
<tr>
<th>Method</th>
<th>Local</th>
<th>Global</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>DTFE</td>
<td>—</td>
</tr>
<tr>
<td>Higher order</td>
<td>NNFE</td>
<td>Kriging, Inverse Distance</td>
</tr>
<tr>
<td></td>
<td>SPH</td>
<td>RBF, Moving Least Squares</td>
</tr>
</tbody>
</table>

Some of the most common interpolation methods are: linear and higher order triangulation based methods (see section 5.4 & 5.5), inverse distance based methods (IDW), moving least squares, radial basis function interpolation (RBF) (appendix 5.C) and Kriging interpolation (section 5.6). Table 5.1 gives an overview of the methods into categories, global & local and linear & higher order. Both RBF and Kriging use predefined kernels to interpolate the field. In RBF the kernel is basis-function that spans the space of all the interpolating functions. On the other hand in Kriging the kernel is the covariance function of a Random Spatial Gaussian process. Since the cosmological density field has evolved from an initial density field that (most probably) has a Gaussian character, we regard the Kriging interpolator as the natural choice for reconstruction of cosmological density fields.

Wiener Filtering is another Bayesian reconstruction approach closely related to Kriging interpolation (see also Zaroubi 2002). However the adopted philosophy behind it is slightly different. Wiener Filtering includes a model for the noise. This implies that the retrieved field is an optimally filtered Gaussian Random Field at various unknown scales. This filter scale depends on the local estimated noise, more noise directly increases amount of smoothing. The disadvantage (or advantage depending on ones goal) will be that nonlinear structures are also substantially diluted or even removed. In section 5.6.4 we will show the technical reason why these nonlinearities had to be filtered out. Therefore Wiener Filtering may have difficulties in dealing with reconstruction the intricacies of the nonlinear structures. Here we specifically aim to include the (quasi-)nonlinear components in the density field, i.e. the Cosmic Web. There are two important consequences of this choice, first we have to be aware that the noise components (Schaap 2007) will be reconstructed as well. Secondly, nonlinear data is typically not well behaved, and we will have to take extra care of nonlinear sites present in the data.

The Delaunay Field Tessellation Estimator (DTFE) was (developed by Schaap & van de Weygaert 2000). The method uses the Delaunay triangulation to reconstruct in a self-adaptive, mass conservative and parameter free way the underlying spatial (density) probability distribution. The DTFE formalism was shown to recover both the hierarchical and anisotropic morphology of the Cosmic Web Schaap (2007); van de Weygaert & Schaap (2007).

Figure 5.2 summarises the intention of this paper. A simple anisotropic Gaussian is irregularly sampled (first column) and reconstructed by the three interpolation methods DTFE (Schaap & van de Weygaert 2000), natural neighbour interpolation (Sibson 1981; Watson 1992) (abbreviated as NNFE) and Kriging (Krige 1951; Matheron 1963) (top row). The middle row provides a comparison between the original contour lines (black) and the reconstructed contours in red. The errors in the reconstruction, measured as absolute deviation from the original, are shown in the bottom row. A first impression is that the higher order interpolation methods provide more accurate reconstructions, and are worthwhile to investigate. Furthermore from table 5.1 we see that the three selected methods fall into the three distinct classes: local & linear (DTFE), local & higher order (NNFE) and global & higher order (Kriging). Using these three methods we will study the properties of the reconstructed galaxy density fields.

A schematic outline of the reconstruction process can be found in Figure 5.3. The data (or point samples) are introduced in section 5.2. Subsequently, we treat the estimation of the
Figure 5.4—The top figure shows the SDSS galaxies in the $X$ and $Y$ coordinates out to a distance of $300h^{-1}\text{Mpc}$. The lower figure shows a $XZ$ slice perpendicular to the $XY$ plane at $Y=135h^{-1}\text{Mpc}$. The corresponding boundaries of both slices are indicated in the other figure, e.g. grew lines in the top figure show the $Y$ limits of the bottom figure and vice versa. Both slices have a thickness $10h^{-1}\text{Mpc}$.

local density for each galaxy in section 5.3. The interpolation methods are explained in section 5.4, 5.5 and 5.6. The post-processing involves filtering of the recovered fields to deal with both the shot-noise and if necessary selection effects. The analysis of the errors is provided in section 5.9. Finally in section 5.11 the SDSS reconstruction is presented. Throughout the paper we focus on a specific case, the reconstruction of the (mock) SDSS density field. We expect that the overall results carry over to other redshift surveys, though the exact limits and scales may differ. The next two Chapters will focus on the cosmography of the reconstructed Large Scale Structure (Chapter 6) and the one-point Probability Density Function (Chapter 7).

5.2 The Data

We have focused on the reconstruction of the density field traced by the spatial distribution of galaxies. The first one is the SDSS galaxy sample itself, the other is a set of galaxy mock catalogues that model the SDSS.
5.2.1 The SDSS galaxy sample

For the (error) analysis we used the main galaxy sample in the North Galactic Cap from the 6th data release of the Sloan Digital Sky Survey (SDSS) (Strauss et al. 2002; York et al. 2000; Stoughton et al. 2002; Adelman-McCarthy et al. 2008). This spectroscopic galaxy sample is almost complete between a Petrosian magnitude limit of $m_r = 14.5$ and $m_r = 17.77$. The SDSS DR6 data release consists of various contiguous regions. We have used the northern strip of the NGC part, the largest contiguous region. The three narrow stripes in the Southern Galactic Cap are not very well suited for a 3 dimensional density reconstruction. The equatorial strip, the other contiguous region at lower declination, is much more narrow than the northern strip, while it has some large gaps in the sky coverage affecting the full 3d reconstruction. One effect is due to fiber collisions where at higher angular densities on the sky some galaxies may have been missed. Although the completeness has some angular variations (Blanton et al. 2003), in this the paper we assume that the completeness does not vary significantly.

The final set of galaxies were selected to lie within a comoving box of $600 h^{-1}\text{Mpc}$. In this box the observer is located at the normalised (divided by $600 h^{-1}\text{Mpc}$) coordinates of $(x, y, z) = (0.5, 0.5, 0.5)$ or $(X, Y, Z)$ of $(300, 300, 0) h^{-1}\text{Mpc}$. The approximate centre of the northern strip is rotated to lie parallel to the y-axis starting at $(x, z) = (0.5, 0.5) = (300, 300) h^{-1}\text{Mpc}$. Within this selected volume there are a total of 311474 magnitude limited galaxies. In Figure 6.1 the galaxies are plotted in the XY-projection. The other plot shows galaxies in XZ-projection (again the observer is situated here at $X = 300 h^{-1}\text{Mpc}$ and $Y = 0 h^{-1}\text{Mpc}$). The density field is reconstructed on a grid of $512^3$, providing a resolution element of $1.17 h^{-1}\text{Mpc}$. 

Figure 5.5—left: A point set and the corresponding Delaunay triangulation is shown in gray. For the central point the natural neighbours are indicated by the blue points with their corresponding delaunay edges. The contiguous voronoi cell/umbrella is shown by the hatched region. right: The black lines show the Voronoi tessellation for points indicated by the diamond shaped points. Inserting the central point (black triangle) gives a new Voronoi cell (dark blue), called the second order Voronoi cell. The gray shaded areas are the overlapping regions with the older Voronoi tessellation.
5.2.2 Selection function

A volume limited sample consists of a subset of galaxies which are homogeneously sampled throughout space. This has the advantage that each point traces in the same way the underlying galaxy density field. A disadvantage of such a scheme is that one discards information from nearby fainter galaxies.

A magnitude limited sample does include this information. On the other hand, one needs to correct for the inhomogeneous selection process. This is achieved by giving extra weight to more distant galaxies, determined by the reciprocal selection probability at the specific distance of each galaxy. The selection function as function of redshift is well fitted by the following equation (Efstathiou & Moody 2001)

\[
N(z) = A z^2 \exp \left( -\frac{z}{z_r} \right). \tag{5.1}
\]

The first term, \( A \), is a normalisation constant, the second term describes the increase in volume, the last term in Eq. 5.1 represents the selection probability. The weights are then given according to

\[
w(z) = \exp \left( \frac{z}{z_r} \right). \tag{5.2}
\]

In principle, the normalisation of the density field can be achieved by modelling the details of the selection function and calculating the appropriate normalisation constant. Alternatively, the nonzero mean can be calculated from the normalised density field. This mean may be subtracted from the reconstructed density. We chose to implement the latter, which is simple and valid as long as the volume is large enough to provide a representative estimate of the mean.

5.2.3 The SDSS mock sample

The Millennium simulation (Springel et al. 2005) was used to construct SDSS like mock catalogues. The semi-analytical galaxy samples (De Lucia & Blaizot 2007; Bower et al. 2006; Bertone et al. 2007) were used to construct our own mock samples. Though these have not been constructed with the same amount of details as for example the Mock-catalogue generated by Blaizot et al. (2005), they are useful as they enable a direct comparison to the full catalogue. Hence we may directly estimate the errors induced by the magnitude selection, peculiar velocities as well as errors caused by the mask.

We generate and define the mock samples on the basis of the following set of conditions;

1. Periodically tiling the Millennium cube to get enough volume.
2. Calculate the redshift w.r.t. to the origin
3. Convert the absolute magnitude to apparent magnitudes using the redshift
4. Select galaxies brighter than \( m_r = 17.77 \).
5. Add the peculiar velocity to the redshift
6. Apply the observational mask of the DR6-NGC

The same coordinate system and box size have been used for the mock galaxies, thus they all reside within a cube of 600 \( h^{-1}\)Mpc.
We will distinguish between a couple of datasets. The full-catalogue where no magnitude selection was done and simply all the Millennium galaxies within the mask were selected. The magnitude limited mock catalogue representing steps (i) through (vi). And a volume limited set out to a redshift of \( z = 0.1 \). Here we removed all the mock galaxies fainter than an absolute magnitude of \( M_r < -20.45 \). These represent roughly the galaxies that are brighter than \( L \). Figure 5.15 presents slices through the three datasets.

### 5.3 Local Density Estimates

The reconstruction of the galaxy density field consists of two essential steps; first the density for each galaxy has to be estimated using a representative local volume or distance. Then the next step is to interpolate the density values to obtain a density field. This should preferably be done in a continuous and mass conserving way.

Local density estimators are for example the distance to the \( n \)th nearest neighbour, the Voronoi volume (Browne 1965; Ord 1978; Okabe et al. 1992; Ebeling & Wiedenmann 1993), the binsize of a kD-tree (Ascasibar & Binney 2005; Sharma & Steinmetz 2006) or a Kernel density estimator like the Epanechnikov kernel (Einasto et al. 2007). Here we use the local DTFE density estimation, as introduced by of Schaap & van de Weygaert (2000), see also Appendix 5.B. It uses the tetrahedra of the Delaunay triangulation to obtain a local density measure (for an extensive overview see van de Weygaert & Schaap 2007). The key point is that the total sum of volumes from the neighbouring tetrahedra belonging to a particle or galaxy, is a fair and self adaptive proxy for the density. The hatched region in Figure (5.5) illustrates this ‘volume’ in two dimensions. This volume was dubbed the contiguous Voronoi cell, and is also referred to in the computational geometry as the umbrella (for obvious reason).

*Note that vertex of Delaunay Triangulation, sample point and galaxy coordinates are the same.
Schaap (2007) showed that the estimation of the density according to the triangulation has several advantages. It self-adaptively resolves density and local shape. Secondly, in combination with a linear interpolation scheme it is also a mass conservative. Using a Sonneira-Peebles particle distribution it was shown that DTFE provides a superior estimate of the one-point distribution function.

In summary, the density value at a certain galaxy position is done according to:

1. construct a Delaunay triangulation (e.g. using the CGAL library)
2. for each galaxy compute the volumes $V_j$ of the $N$ neighbouring tetrahedra $T_j$
3. determine the density at site as the inverse of the sum of the volumes $V_j$.

\[
\hat{\rho}_i = \frac{4 \ w_i}{\sum_{j=0}^{n} V_j}
\]  

(5.3)

the factor four takes account of the fact that each vertex is counted four times in 3D, and $w_i$ are the weights according to equation (2).

We note that instead of calculating the density particle by particle, it is more efficient to loop over the tetrahedra and increment the density for all of the four vertices of the tetrahedron. The weights $w_i$ are given by the selection function.

### 5.3.1 Shot noise errors

A local density estimator such as the contiguous Voronoi cell has the advantage that it is very sensitive to the signal. However they also have the property to pick up the shot noise present in the data. The statistical properties of the density estimator are described in Schaap (2007). An important result they found was that for a (3D) Poisson point process with intensity $\lambda = 1$ the probability distribution of the above estimator is given by:

\[
p(\hat{\lambda}) = \frac{1944}{5} \hat{\lambda}^{-8} e^{-6/\hat{\lambda}}
\]  

(5.4)

The estimator is unbiased as the mean of $p(\hat{\lambda})$ is equal to one. The variance $\sigma^2 = \frac{1}{3}$ gives an estimate of the estimator error, and it is roughly 57%. The distribution is non-Gaussian with a large tail extending to high density values (Fig. 5.6).

To get a visual impression of the shot noise present in the density estimate, we generate a Poisson point sample of an anisotropic Gaussian distribution of 4500 points, see Figure 5.8. This random point sample has shot noise and provides a more meaningful and realistic example of what may be expected for the galaxy distribution. The densities were estimated as described above and linearly interpolated (see section 5.4) to obtain a reconstructed field. The resulting density field provides a good illustration of the shot noise component. The reconstructed blue contour levels are highly irregular (top-middle panel) compared to original contours (top-left). The bottom-left panel shows a profile through the reconstructed field (orange line), it gives a visual impression of the scatter in the reconstruction. Note that the sharp peaks give rise to high density tail in eq. 5.4.
5.3.2 Centroidal Voronoi Tessellation

The high density tail can be suppressed or regularised by using the centroidal Voronoi tessellation (CVT) (see Lloyd 1982; Browne 2007). Calculating the CVT can be done by iteratively by calculating the centre of mass of the Voronoi cells and displacing the particles to that position. Effectively particles repel each other. Eventually the resulting Voronoi diagram converges and attains a regular shape (see Fig. 5.6), where the centre of mass is at the same position as the particle. This final stage is called the Centroidal Voronoi Diagram. Here we shall analyse the iterated steps before the particles have fully converged.

We have applied the above iterations (these iterations are referred to as Lloyd iterations) to a Poisson point process. Figure 5.6 shows the density probability distributions after 0, 2, 4, 6, 8 and 10 Lloyd iterations. The tail of the distribution is almost absent after 4 iterations, providing a narrower and more symmetric distribution. The estimator also remains unbiased as the mean stays approximately unity. We conclude that using the CVT the estimated error gets smaller and less skewed.

The same procedure was applied for the test case in Figure 5.8. After 5 Lloyd iteration particles are much more regular (top-right) and this is also reflected in the contour lines. The bottom-left panel shows a profile in the y-direction at x=0.5. The black curve is the original profile, orange is the DTFE interpolated field and the blue line is the reconstructed profile after 5 Lloyd iterations. It may be obvious from the profiles and the contours that the shot-noise has been greatly reduced with the CVT procedure.

To see how the average trend is affected we have smoothed the field using a Gaussian filter ($R_f = 0.05$). The middle panels shows the smoothed original (left), the smoothed DTFE (middle) and the smoothed field after 5 Lloyd iterations (right). From the panels we may observe that the anisotropy in the xy direction has been affected by the Lloyd iterations. To investigate this we plot in the bottom row the profile along the y-axis at x=0.5 (bottom-middle) and a profile in x-direction at y=0.5 (bottom=right). From the profiles and the contours we can see that contours with 5 Lloyd iterations are worse at “larger scales”. The deviations are higher as a part of the anisotropy has been lost. We may conclude that only at small scales (typically inter-particle distances) the CVT improves the reconstruction, but should be avoided if the density field is filtered at larger scales.

5.4 Delaunay Tessellation Field Estimator

The most straightforward way to interpolate and/or reconstruct a density field is by linearly interpolating between neighbouring data points. This linear reconstructed field is continuous everywhere and the first derivative remains constant inside the tetrahedra. However at the boundaries between tetrahedra the derivative is discontinuous.

One of the key aspects of DTFE (Schaap & van de Weygaert 2000; Schaap 2007; van de Weygaert & Schaap 2007) is the choice of the local neighbourhood that can be considered as a natural higher dimensional extension of the one-dimensional ordering of points. First note that the one-dimensional neighbourhood divides the x-axis into natural bins. These bins are defined by the (two) bisectors of a data-point $d$ with the neighbour to the left and the other neighbour to the right. All the space in between the two bisectors is closer to point $d$ than to any of the neighbours. This concept can be generalised to higher dimensions and is called the Voronoi Diagram (see Figure 5.5).

Just as in the one-dimensional case the reciprocal data structure defines a neighbourhood function: the Delaunay triangulation. The Delaunay triangulation has some unique geometric properties making it a natural choice for the neighbourhood function (Bernardeau & van de
Figure 5.7— The same anisotropic Gaussian distribution (as in Figure 5.2) is Poisson sampled with 4500 points (top-left panel). Top-middle gives respectively the DTFE reconstruction and top-right: the DTFE reconstruction after the particles were displaced using 5 Lloyd iterations. The middle row shows the corresponding contour plots after Gaussian filtering ($R_f = 0.05$). The bottom-left gives profiles at $x = 0.5$ for through the original (black), DTFE (orange) and DTFE+LLoyd5 (blue). Using the same colours the bottom-middle & bottom-right shows the profiles through the filtered reconstruction, at $x = 0.5$ and $y = 0.5$. 
Figure 5.8—The density reconstruction for the three different methods from the distant observer SDSS mock catalogue. With top-left the N-body particles, top-right DTFE density field reconstruction. Bottom left and right are respectively the NNFE and lognormal Kriging reconstructions.

Weygaert (1996). Once it has been constructed (see Aurenhammer 1991) linear interpolation is straightforward, each Delaunay triangle (2D) defines a unique interpolating 3-dimensional plane. Together these triangles form a interpolating piecewise linear surface.

We will now briefly describe the DTFE formalism; given a point $\mathbf{r}$ located in a tetrahedron $T$ with vertex coordinates $r_0, r_1, r_2, r_3$. The estimate of $f$ at a point $\mathbf{r}$ is given by the following relationship;

$$f(\mathbf{r}) = f(r_0) + \nabla f \bigg|_j \cdot (\mathbf{r} - r_0).$$

(5.5)

The components of the four dimensional gradient, $\nabla f|_i$, are found by solving the three linear equations.

The Delaunay triangulation in this work has been computed using Computational Geometry Algorithms Library (CGAL).
sets of equations.

\[
\begin{align*}
    f(r_1) &= f(r_0) + \frac{\nabla f}{\nabla j} \cdot (r_1 - r_0) \\
    f(r_2) &= f(r_0) + \frac{\nabla f}{\nabla j} \cdot (r_2 - r_0) \\
    f(r_3) &= f(r_0) + \frac{\nabla f}{\nabla j} \cdot (r_3 - r_0).
\end{align*}
\]

The density field defined on a grid may be obtained by using the interpolated value for each voxel location. In the appendix 5.B we provide a more detailed outline of the method.

An impression of a DTFE interpolated field in the cosmological context is presented in Figure 5.8. The uppermost left panel shows a slice through the millennium box. The galaxies represent a selection according to a distant observer approximation (i.e. lines of sight that are parallel) with the magnitude limit of SDSS. The panel to the right shows the DTFE field, it recovers both the fine structures at close proximity and adapts itself to larger scales at greater distances in a parameter free way. It also shows a disadvantage of the linear interpolation scheme, linear artifacts are produced when one crosses from one triangle into another triangle. These are especially noticeable when the data points are sparse. We must note that these wings are not significant in mass, but arise when one makes a lower dimensional (1 or 2) cut through the data.

5.5 Natural Neighbour Field Estimator

The DTFE is a piecewise linear interpolation (C⁰) method. In a sense it is a linear version of a larger class of tessellation based interpolation methods, Natural Neighbour interpolation (Sibson 1981; Watson 1992; Braun & Sambridge 1995) is the most well known higher order tessellation based method (for more details see van de Weygaert & Schaap 2007). The interpolation is continuously differentiable with the exception of the sample points where the derivative is discontinuous. The natural neighbours of a vertex are defined as all the points connected to that vertex according to the Delaunay triangulation (see 5.5). The name is derived from the fact that it uses the data from the natural neighbours to obtain an interpolant. The key idea is to find a weight for each neighbour such that \( f \) is interpolated in a continuous manner. In combination with the density estimator of section 5.3 we obtain a reconstruction method, which we call the Natural Neighbour Field Estimator (NNFE).

To determine the interpolation weights \( \lambda \) one may consider the location, \( \tilde{r}_i \), as a virtual point to be inserted in a Voronoi tessellation. In this updated triangulation we use the Voronoi cell of the virtual point \( C_\tilde{r} \) to obtain the weights. If one would insert the virtual point in the existing Voronoi Tessellation (black lines) then a new cell forms, shown in Figure 5.5 as the cell encircled by the blue outline. We distinguish two kind of volumes, the volume of cell \( V_{\tilde{r}} \) and of the volumes of the second order Voronoi cells \( V_i(\tilde{r}) \). These cells are indicated by a different gray level in Figure 5.5. Note that each second order Voronoi cell overlaps uniquely with one of the Voronoi cells of the original Voronoi Tessellation. The ratio between the volume \( V_i \) and the volume \( V_{\tilde{r}} \) provides a unique weight for each of the natural neighbours. Since the complete cell \( C_\tilde{r} \) is decomposed into second order Voronoi cells \( C_i \), the sum of these ratios always sum up to one

\[
\lambda(\tilde{r}; r_i) = \frac{V_i(\tilde{r})}{V_{\tilde{r}}(\tilde{r})}
\]

\[
\sum_{i=0}^{N} \lambda(\tilde{r}; r_i) = 1,
\]
this latter property is called partition of unity. Accordingly an interpolation can be provided by the following equation

\[
\hat{f}(\mathbf{r}) = \sum_{i=0}^{N} \lambda(\mathbf{r}; \mathbf{r}_i) f(\mathbf{r}_i).
\] (5.9)

If one places the interpolation point closer and closer to a vertex \( v \) (galaxy), the overlap with the Voronoi cell belonging to this vertex will steadily increase. The interpolated value will thus be more and more weighted according to the datum of that vertex. When the location finally coincides with one of the vertices, the new second order Voronoi cell is identical to the old Voronoi cell at this vertex. The interpolated value is then the same as the data value at that vertex (eq. 5.9). Natural neighbour interpolation reproduces the known data values, an important requirement for an interpolation method. At the position of the vertices the derivative of the interpolant is discontinuous. Moving the interpolation point in slightly different directions introduces different sets of natural neighbours.

In one dimension DTFE and NNFE are exactly the same. When the data-points are given on a regular grid, the NNFE reduces to the more familiar bi-linear (2d) or trilinear (3d) interpolation schemes. The NNFE method code was implemented by Eldering et al. (2006) and it is a 3 dimensional adaption of the 2d version available in the CGAL library. In Figure 5.8 we show a result of natural neighbour interpolation. Some of the peaks in the regions with sparse sampling appear somewhat anisotropic, relating to the discontinuous derivative at a vertex point as described above. The overall resulting field is nice and smooth without all the artifacts that the DTFE method shows.

### 5.6 Kriging Interpolation

A drawback of the DTFE and the NNFE methods is that both do not take into account the existing spatial correlations present in the cosmological density field. By basing itself on the covariance function of the density field, Kriging naturally includes these global correlations. The method was named by Matheron (1963) after D. G. Krige who started the development of the method (see Cressie 1990, for a historical overview). The interpolator has the property that it is a best linear unbiased estimator (Cressie 1988, 1993, see also appendix 5.D). Most examples stem from the field of geostatistics where it was developed. Examples are the reconstruction of gold, ore and oil fields, altitude maps, etc., where the measurements were taken at spatially scattered points.

Though Gaussian Random Fields and irregular spaced data are quite common in astronomy, the Kriging method is relatively unknown. A notable example is an application by Alfaro et al. (1991), who used it to retrieve the structural properties of the Galactic disk. This absence may be partly due to the fact that very often the data has a strong noise component. In such cases Wiener filtering, a method closely related to Kriging (see Rybicki & Press 1992), is the optimal method for recovering the underlying Gaussian Trend or Field. As earlier mentioned our approach differs from the Wiener filtering, because we explicitly want the filtering process to be the last step.

Another reason why the Kriging method is not that well known in astrophysics while it is one of the main tools for similar reconstruction purposes in geophysics is that both fields use entirely different formulations and nomenclature. One particular example concerns the use of the correlation function or the variogram. For example, the Unbiased Minimum Variance estimator, developed by Zaroubi (2002) is an alternative derivation of the Kriging equations stemming the Wiener filter formalism. Another example of such a different naming convention is that Kriging solution can be considered as the mean field in Constrained Random Field technique as formulated by Bertschinger (1987)(also see Hoffman & Ribak 1991; Rybicki &
Press 1992). In geostatistics this is called Conditional Simulation. In this context we may also regard the lognormal Kriging method used here, as the mean field reconstruction of constrained lognormal random fields described by Sheth (1995). In order not to confuse terms and definitions in the following we will keep the formulation of the Kriging interpolation method below in accordance with the geostatistical literature where it was developed.

5.6.1 The Kriging method

Interpolation can be viewed as a weighted linear combination of nearby known data points $f(r_i)$;

$$\hat{f}(\mathbf{r}) = \sum_{i=0}^{N} \lambda_i f(r_i)$$  \hspace{1cm} (5.10)

Here $\hat{f}$ is the estimated value at the location $\mathbf{r}$.

The main idea of Kriging is to calculate the values of weights $\lambda_i$ that minimise the error with respect to the data according to the mean square variation,

$$E(||\hat{f}(\mathbf{r}) - f(\mathbf{r})||^2).$$  \hspace{1cm} (5.11)

In Appendix 5.D we show a derivation of the Kriging linear system. Also we discuss the statistical properties of the estimator, the variance in the estimate and the relation it has with constrained random fields.

In geostatistics the spatial dependence of random field, $f(\mathbf{r})$, may be characterised with a variogram. The variogram is the mean square variation as function of distance defined as

$$2\gamma(\mathbf{r}_1, \mathbf{r}_2) = E(||f(\mathbf{r}_1) - f(\mathbf{r}_2)||^2).$$  \hspace{1cm} (5.12)

For a stationary random field this reduces to;

$$2\gamma(h) = E(||f(\mathbf{r}) - f(\mathbf{r} + \mathbf{h})||^2).$$  \hspace{1cm} (5.13)

The variogram is related to the covariance function $c(h)$,

$$c(h) = E(f(\mathbf{r}_1)f(\mathbf{r}_2)) = c(0) - \gamma(h).$$  \hspace{1cm} (5.14)

There are a variety of variogram models (see Cressie 1993, for a detailed description). We use the exponential form

$$\gamma(h) = \sigma_0(1 - \exp((h/h_0)\phi)),$$  \hspace{1cm} (5.15)

which provides a good fit for our measured variograms (see section 5.6.3).

The simple Kriging equations are given by

$$\mathbf{C}(\mathbf{r}_i, \mathbf{r}_j)\lambda_j = \mathbf{c}(\mathbf{r}_i, \mathbf{r}),$$  \hspace{1cm} (5.16)

where the matrix elements of $\mathbf{C}$ are given by

$$\mathbf{C}(\mathbf{r}_i, \mathbf{r}_j) = E(f(\mathbf{r}_1)f(\mathbf{r}_j))$$  \hspace{1cm} (5.17)

and vector elements are given by

$$\mathbf{c}(\mathbf{r}_i, \mathbf{r}) = E(f(\mathbf{r}_i)f(\mathbf{r}))$$  \hspace{1cm} (5.18)
Figure 5.9— The measured variogram from top to bottom correspond to the unsmoothed field (red), and the filtered fields at scales of $1h^{-1}\text{Mpc}$ Gaussian filtered, $3h^{-1}\text{Mpc}$, $6h^{-1}\text{Mpc}$ and $10h^{-1}\text{Mpc}$ (black). The fitted variogram models according to eq. 5.15 are shown in orange.

Usually the covariance function only depends on distance, $d = |\mathbf{r}_i - \mathbf{r}_j|$. One can solve this linear system of $N$ equations for the $N$ unknown weights, $\lambda_i$. These weights have to be specifically computed for each interpolation site, $\mathbf{r}_i$ (though the matrix has to be inverted only once). Subsequently one can compute the interpolated field values $\widehat{f}$ from equation 5.10.

There are many variants on the Kriging equations. We deal with ordinary Kriging when the mean of the field is unknown. It uses an extra Lagrange multiplier for the unknown mean. We speak of Universal Kriging is when the mean is not constant. In this case we may need to include a trend function. This involves the introduction of several Lagrange multipliers, each multiplier corresponding to a co-factor of a polynomial trend function.

5.6.2 Practical Issues

Two main practical problems exist using the Kriging interpolation method for galaxy density fields. The first is that the underlying field we would like to reconstruct is not Gaussian. It has been altered by nonlinear gravitational growth. Gravity empties out large underdense regions, while at same time overdense clumps collapse to provide very high density values. The one-point probability density distribution is thus characterised by a peak at low densities and a tail extending to very high densities. We will show that especially the high density tail leads to a fundamental problem for higher order global interpolation schemes.

The second issue is a more practical one. Kriging, like Wiener filtering, is a linear global estimator. This requires the solution of a large linear system, corresponding to the inversion of a large covariance matrix $C_{ij}$. This matrix is of the size of the number of galaxies. Typical sizes for modern redshift surveys are of the order of hundred thousand to a million galaxies or more. Inversion of such matrices is either too slow or simply not possible. Practical solutions are to reduce the size of the matrix or alternatively use numerical approximations to obtain the inverse matrix (as explored in Kitaura & Enßlin 2008).

In the upcoming section we will address both problems and propose a practical solution that for most part deals with both problems. While we focus specifically on Kriging, preliminary tests show that the results carry over to other global methods like Radial Basis Function Interpolation (Appendix 5.C), Inverse Distance Weighted Interpolation and possibly also to the Wiener Filtering Method.

5.6.3 Variogram estimation

Kriging requires a functional form of the covariance function, which is related to the variogram as in eq. (5.14). To estimate the variogram we use the estimator as provided by Cressie
Table 5.2— Fitted Parameters of the cosmological simulation that were used for the Mock Catalogues.

<table>
<thead>
<tr>
<th>( R_f (h^{-1}\text{Mpc}) )</th>
<th>( \sigma_0 )</th>
<th>( h_0 )</th>
<th>( p )</th>
<th>( R_{\text{max}} (h^{-1}\text{Mpc}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2.78</td>
<td>8.4</td>
<td>1.2</td>
<td>200</td>
</tr>
<tr>
<td>1</td>
<td>2.65</td>
<td>9.6</td>
<td>1.5</td>
<td>200</td>
</tr>
<tr>
<td>3</td>
<td>1.85</td>
<td>13.5</td>
<td>1.72</td>
<td>300</td>
</tr>
<tr>
<td>6</td>
<td>1.0</td>
<td>18.6</td>
<td>1.9</td>
<td>400</td>
</tr>
<tr>
<td>10</td>
<td>0.45</td>
<td>24.0</td>
<td>2.1</td>
<td>500</td>
</tr>
</tbody>
</table>

\[ \tilde{\gamma}(h) = \left( \frac{1}{N_p} \sum_{i=0}^{N_p} \sqrt{|f(r) - f(r + h)|^4} \right) \]  

(5.19)

Nonlinear clustering of the galaxies also affects the estimation of the variogram. Therefore we do not estimate the variogram at the data points themselves, but rather use a set of randomly chosen locations. At these sites the density is provided by the logarithmic value of the DTFE-interpolated field. Then using Eq. 5.19 the variogram the variogram is estimated.

The variogram models were estimated from the Millennium mock galaxies using distant observers approximation (see figure 8). Figure 5.9 shows the measurement for the case of no smoothing (red) and the smoothed fields with respectively a scale of 1, 3, 6, 10 \( h^{-1}\text{Mpc.} \) The fitted variogram models according to equation 5.15 are superimposed with parameters as given in Table 5.2. It turns out that the interpolated values do not depend very much on the precise choice of the variogram parameters. This is a general property of Kriging (see Press 2007), and is mainly of concern for the estimated error. Henceforth we will use the covariance kernel corresponding to a \( R_f = 0 \ h^{-1}\text{Mpc} \) (i.e. with parameters given by the first row in Table 5.2) for all of the Kriging reconstructions.

5.6.4 Nonlinear Kriging

Kriging is a Gaussian regression method, it assumes that the data is derived from a spatial Gaussian random process. Even if the data is not Gaussian it has been noted that the Kriging estimator fairs surprisingly well. In cases where it does fail, a common approach is to use a certain mapping, \( \phi \), that transforms the density values to a new set of density values. And \( \phi \) has been chosen such that this new set of density values has a Gaussian one-point probability density distribution (Hawkins & Cressie 1984; Matheron 1976; Cressie 1993). Subsequently, use these transformed density values, \( \phi(f) \), to interpolate a density field using the standard Kriging procedure. Following the interpolation, the interpolated values are transformed back using the inverse mapping, \( \phi^{-1} \), providing the final answer. Note that such a interpolated field is no longer a linear combination of the data \( f \), but a linear combination of transformed data \( \phi(f) \). In other words we have replaced Eq. 5.10 by;

\[ \widehat{f}(\mathbf{r}) = \phi^{-1} \left( \sum_{i=0}^{N} \lambda_i \phi(f(\mathbf{r}_i)) \right). \]  

(5.20)

When the transformation, \( \phi \) consists of a linear combination of Hermite polynomials, we speak of disjunctive Kriging. Similar to the Edgeworth expansion used in perturbation analysis
(Juszkiewicz et al. 1995; Kofman et al. 1994; Bernardeau & Kofman 1995), it exploits the fact that Hermite polynomials are orthogonal with respect to a Gaussian distribution.

Another form of non-linear Kriging method is lognormal Kriging. The lognormal distribution given by

\[ P_{LN} = \frac{1}{\sqrt{2\pi S^2}} \exp \left( -\frac{[\ln(1 + \delta) + S^2/2]^2}{2S^2} \right) \frac{1}{1 + \delta} \tag{5.21} \]

where \( S = \ln(1 + \sigma) \) and \( \sigma^2 \) is the variance of the density field, see Coles & Jones (1991). This procedure uses the logarithm to transform the data, i.e. \( \phi = \log(f) \) and can only be used for positive valued fields. Since the density \( \rho = \bar{\rho}(1 + \delta) \) is always positive a lognormal approach is valid. The final interpolation values are then obtained by taking inverse transformation, the exponent of the interpolated data values;

\[ \hat{f}(r) = \exp \left( \sum_{i=0}^{N} \lambda_i \log(f(r_i)) \right) \tag{5.22} \]

A strong justification for using lognormal Kriging is that the evolved cosmological density field can be characterised by a lognormal distribution (see Coles & Jones 1991; Kofman et al. 1994; Sheth 1995; Kayo et al. 2001). In Chapter 7 we shall show that the galaxies density distribution is indeed very well modelled by a lognormal distribution at scales equal or larger than \( 3h^{-1}\text{Mpc} \).

5.6.5 Peak-Correlation Overshoot & Gibbs oscillations

Applying the linear Kriging model on galaxy or N-body simulation data gives strongly distorted interpolated fields. Kriging is a global method, hence nearby and or even far away sites may couple to highly nonlinear objects like clusters. These nonlinearities have densities
that are three to four magnitudes larger than local field. Moreover, many of the data points themselves are located in such sites, which leads to a complete overestimation of the inferred density values in the neighbourhoods of peaks. We call this peak-correlation overshoot.

To illustrate the peak-correlation overshoot we construct an easily interpretable one dimensional case of the previous interpolation problem. We sample two smoothly varying cosine functions with 55 random points between 0 and 2. Subsequently 5 points are removed between $x=[1.,1.5]$ and reallocated at $x \approx 0.7$. These points are given anomalously higher density values $f_{\text{peak}}$, representing a nonlinear peak in an otherwise smooth and well behaved density field. A one dimensional illustration of the constructed function can be seen in Figure 5.10.

In this case the normal Kriging method does not yield any interpolating solution. One can give a certain amount of variance at zero distance. This is called nugget-effect: it may be thought of as attributing to each point a certain uncertainty so that interpolation does not necessarily need to pass trough each data point. In a sense it turns the interpolation method into a fitting routine (see Press 2007, for a deeper discussion).

Adding a variance at zero lag of a few percentage the variance provides an ‘interpolated’ solution for both the normal Kriging as well as the lognormal Kriging. The two reconstructions are plotted in Fig. 5.10. The peak-correlation overshoot is visible as the oscillatory behaviour around the peak, essentially related to Gibbs oscillations. Also notable is that the oscillation are much less pronounced in the lognormal version, however they are still present. Considering that the (nonlinear) dynamic range of the density contrast of the density field at for example a scale of $1h^{-1}\text{Mpc}$ is 3 orders of magnitude, $\Delta \approx 100$. The latter implies that even taking the logarithm will not suffice to kill these oscillations in case of the galaxy density field.

We may observe that the oscillations vanishes at greater distances from the peak. This is related to the typical correlation scale, which in this example is roughly equal to one wavelength (of the underlying cosine function!). Since the Kriging weights $\lambda_i$ are roughly proportional to the correlation strength. This implies that at a typical distance of one wavelength from the peak the Kriging weights $\lambda_{\text{peak}}$ go to zero. Subsequently, also the term $\lambda_{\text{peak}}f_{\text{peak}}$ in the summation of eq. 5.10 vanishes. In a sense if one is sufficiently far away from the peak the interpolation will not sense the presence of the the peak.

This example can be taken to an extreme limit, where the data is completely Gaussian except for one nonlinear outlier in the field, where the density has an extreme value $f_{\text{peak}}$. Many interpolation methods, including Kriging, will fail to provide a good reconstruction for this example. As pointed out by Hawkins & Cressie (1984), eq. (5.10) is not robust against outliers. Suppose a point at large distance has a very small weight $\lambda$. One may always increase $f_{\text{peak}}$ to such high values that peak term $\lambda_{\text{peak}}f_{\text{peak}}$ in eq. (5.10) is non-zero, and thus corrupting the interpolated value out to very large distances.

Hawkins & Cressie (1984) suggest a robust implementation of the Kriging method. They replace eq. (5.10), which is a weighted average, by a weighted median. In our 1d-example we have applied this second method. The resulting light-blue curve in Figure 5.10 shows that it indeed produces a non-oscillatory approximation. However it is based on a median and it does not interpolate in a continuous manner. Instead it jumps in between two consecutive points. The staircase appearance of the curve can be clearly seen at the location where the data points are sparse ($x \approx 1.4$).

Though this robust method may not be of practical use here, it does provide some insight in how to get rid of the oscillations. The median simply discards the outliers in the estimate. A similar result can be achieved by using adapted local stencils, restricted to only the well-

\footnote{A stencil is a region or a set of points that provides the support of the approximating function.}
behaved data points. This leads to so called essentially non-oscillatory interpolation methods or ENO-schemes. These are piecewise polynomial schemes on local stencils (Harten & Osher 1987). These stencils can be expanded to general neighbourhoods. An example application to unstructured meshes can be found in Abgrall (1994). Localised implementations of the Kriging method are often discussed to reduce the computational costs. Here we will also use it to keep the outliers as local as possible, while retaining the benefits of a global interpolation scheme.

5.6.6 Natural Lognormal Kriging

Local versions of Kriging interpolation schemes were introduced to speed up the algorithm. This is allowed when nearby points carry the largest weight and are of most influence. The far larger number of more remote points may then be ignored. Two common choices to reduce the set of points exist:

- use only the points within some given distance
- take a fixed amount of $n$ nearest neighbours.

The $n$ Kriging equations are then solved for the reduced set of points. Obviously a general trade-off exists between accuracy and speed. We choose to follow the brute force option; “use as many as we can”. Suggestions for a “as many as we need” solution are provided in section 5.12.

The local set of points that we chose are the natural neighbours of the tetrahedron in which a point is located, which we call the tetrahedral natural neighbourhood. These are the union of all the natural neighbours belonging to the four vertices of the tetrahedron in which a point is located. Figure 5.11 displays the tetrahedral natural neighbourhood in a 2 dimensional case.

We regard this neighbourhood superior to either of the above 2 choices. The distance based neighbourhood relies on one particular scale to represent the data and ignores any possible clustering in the data. At high densities the typical distance would be too large. It would therefore select too many points, and we would again end up with a large linear system. It also fails at low density where not enough points will be selected.

The other option of choosing the $n$-nearest neighbourhood is locally adaptive. However it has the disadvantage of over-representing clustered data points. Instead our choice draws upon the self-adaptiveness of the Delaunay triangulation, and does not suffer from such adverse effects.

We found that on average in 3 dimensions the tetrahedral natural neighbourhood contains approximately 57 particles. One can extend the neighbourhood by adding a third or more layers around it. An additional third layer would yield an average of 284 neighbours. This makes the extension rather slow, while the overall difference are not that large.

Localising the method has one major downside, it may become discontinuous. New particles and data are introduced when the interpolation crosses a boundary of a tetrahedron. We will assume that discontinuities mostly happen in case of peak-correlation overshoot. Here we actually prefer a local discontinuity (in the gradient or higher derivatives). This is also the reason that the linear DTFE-method is quite robust, outliers and other discontinuities are strictly of local influence. In the one-dimensional case we will show that the combination of both transforming the data as well as using only local data, produces a method, Natural Lognormal Kriging (see Appendix 5.E) that is far more robust against outliers. The term natural derives from adopting the natural neighbours as the local neighbourhood. Another benefit is that the computational costs are greatly reduced by only considering local neighbourhoods.

Focusing again on the 1d example, the localised lognormal Kriging is shown in the left panel of Figure 5.10). We choose six neighbours, three to the left and three to right for each sampling point. It greatly reduces the overshoot problem, most of the induced oscillations
Figure 5.11—Tetrahedral natural neighbourhood: for the same point sample as in Fig. 5.5, we indicate the tetrahedral natural neighbourhood of the blue cross by the red outlining. The dashed blue triangle with the blue points (astrixes) indicates the triangle where the sample point is located, together with the gray points (astrixes) they form the subset of points in this neighbourhood.

Figure 5.12—Benefit of the higher order interpolation methods. The figure compares linear to higher order interpolation in the case of sparsely sampled data points. The black curve shows the original double cosine function. The orange and red curve are respectively the linear DTFE and Natural Lognormal Kriging approximating functions.

are gone, because the information of the peak is locally contained. Within the natural neighbourhood the interpolation still shows some minor distortions. For comparison we have also plotted the DTFE interpolated function. As mentioned before, the local linear interpolation performs better around the discontinuity.

The benefits of higher order schemes are most obvious when data points are sparse. We plot in Figure 5.12 the region where we have removed 5 sampling points (around $x \approx 1.3$). Here higher order methods produce much better results than the linear interpolation. The reason is that higher order methods are able to approximate the higher order terms of the underlying function (see Franke 1982). In the context of cosmological density fields this will mean that the field should be more smoothly reconstructed in regions where the sampling is
sparse, the underdense regions.

5.7 Computational Requirements

A first factor of importance in evaluating the performance of the DTFE, NNFE and Natural Lognormal Kriging methods concerns their computational demand. This involves both memory and cpu requirements.

The major component of the memory consumption is the construction of the Delaunay triangulation. As rule of thumb in CGAL the memory allocation increases by one Gigabyte for each additional million particles.

The CPU requirements are foremost determined by the construction of the Delaunay triangulation and the interpolation step. The latter is very much dependent on the grid size and the method. In case of the DTFE this is an order of magnitude faster than NNFE and the Natural Lognormal Kriging is again an order of magnitude slower than NNFE. This is a reflection of the increasing neighbourhood that causes the latter two methods to be much more CPU intensive. As an illustrative example the SDSS (mock) reconstruction on a $512^3$ grid (cf. figure 5.29, 5.30). For DTFE it took approximately half an hour on a single processor, NNFE used half a day of computing time. The Kriging method took several days using multiple processors (the grid was divided up into four different slices).

5.8 Qualitative Density Comparison

The underlying density field for the galaxies is not known, so one can not give direct errors of the reconstructed density field. However, we can provide a measure of consistency by comparing the reconstructed density fields from the mock catalogues with the density fields generated from the full galaxy catalogue. First we will give a qualitative comparison of the visual appearance of the density fields. Followed by a quantitative analysis of the errors, correlation and topology in the reconstruction.

In figure 5.13 we show for the three reconstruction methods the contour maps derived from the full millennium galaxy catalogue. In the next figure 5.14 we show at the same coordinates the reconstructed density fields for a volume limited mock catalogue. To be able to differentiate between underdense regions and overdense regions the underdense regions are filled in with coloured contour, whereas the overdense regions are represent by the black contours. The contour levels in all the plots were kept at the same densities.

5.8.1 Maps of the density field

The first impression is that the DTFE density field shows quite some sharp artifacts (see also appendix 5.F). This is especially clear in the mock reconstructed field of in figure 5.14. These artifacts are leftovers from the linear interpolation method. When two neighbouring grid cells are located in different triangles, artifacts can arise. The field appears to be discontinuous, though a sampling at finer scales would show that this is not the case. When the point distribution is much sparser than the size of the grid cells such artifacts become apparent. This is the reason why they are more pronounced in the mock sample (Figure 5.14) than in the full sample of Figure 5.13.

The two higher order schemes do not show such artifacts. Whereas the overall appearance of both the NNFE and the Kriging fields are strikingly similar. Both methods are able to interpolate smoothly inside voids as each interpolation site can use information from more distant vertices. If we compare the 4 vertices (tetrahedron) that DTFE used for interpolation, to the average of 17 natural neighbours that NNFE uses and the 57 vertices used by Kriging. We see that higher order methods use approximately an order of magnitude more information.
Figure 5.13 – The figure is a comparison between the three methods for the full millennium sample using the SDSS-DR6 and redshift space distortions. From top to bottom the three panels are the DTFE (top), NNFE (middle) and Kriging (bottom) density reconstruction. The coloured contour levels represent the underdense regions, they are resp. [0.001, 0.002, 0.005, 0.01, 0.02, 0.03, 0.05, 0.07, 0.1, 0.2, 0.3, 0.6, 0.7, 0.8, 1]. The white areas are the overdense regions and the black contour lines represent a density contrast of [1, 3, 10].
Figure 5.14— Similar as in figure 5.13 using the same contour levels we plot the DTFE, NNFE and Natural Lognormal Kriging reconstructed fields. Each was reconstructed from the volume limited galaxy catalogue with redshift space distortions.
Figure 5.15– The figure shows a direct comparison between the density reconstruction from the full Millennium catalogue, a magnitude limited catalogue and a volume limited catalogue. For the full catalogue the galaxies of a $6.0 \, h^{-1}\text{Mpc}$ thick slice are plotted. Galaxies from magnitude and volume limited catalogues were selected using a $12.0 \, h^{-1}\text{Mpc}$ slice. The contour lines are from the NNFE density reconstruction and are given for each reconstruction at a density contrast value of 0.25 (blue) and 1.0 (orange).
Figure 5.16— The figures show a line of sight density profile through the density cube. The profile is along the Y-axis at $X=300h^{-1}\text{Mpc}$ and $Z=300h^{-1}\text{Mpc}$, similar coordinates as in figure 13-15. From top to bottom we use the DTFE, NNFE and Natural Lognormal Kriging reconstruction techniques. The black line represents the density profile through the full galaxy sample ($R_f = 1.0h^{-1}\text{Mpc}$), whereas the red line is corresponding profile through the reconstructed density field from the mock sample. The gray (full) and the orange (mock) lines are the profiles at the same location using the $10.0h^{-1}\text{Mpc}$ filtered fields.
Figure 5.17— For the profiles in figure 5.16 the errors $\epsilon_1$ (see equation 5.24) are plotted in at the same coordinates. For comparison we also show the DTFE profile at the top, below this plot we show the DTFE errors, NNFE and Natural lognormal Kriging errors in the lowest panel. The green line corresponds to the errors of the unsmoothed profiles, the errors of the $10.1h^{-1}\text{Mpc}$ filtered field is shown in purple.

One of the crucial benefits of DTFE is that it is able to pick up anisotropic features, like walls and filaments (see Schaap 2007; van de Weygaert & Schaap 2007). One may therefore wonder whether higher order methods (NNFE and Kriging) also have retained this ability. From Figure 5.14 we may see that the overdense regions (black and white shaded) in NNFE and the Kriging fields are much smoother than in the DTFE field. Though if we look at Figure 5.13 only some small details in the overdense contours differ. Since the latter field was sampled by many more points, we expect here no major differences. And in general we do not observe a systematic break up or disappearance of filamentary structures. We may conclude that both higher order methods are able reconstruct the filamentary patterns of the Cosmic Web. Plus higher order methods have the benefit that the contours at sparser sampling are less noisy than the DTFE reconstruction.

5.8.2 Density Profiles

For a more detailed assessment we show in Figure 5.16 a profile of the density field through the box. The three panels correspond to the reconstruction for the three different methods. The full density field (black line) is plotted together with the reconstruction from the mock catalogue (red). We also show the profiles of the fields that were smoothed with a Gaussian at a filtering scale of $R_f = 10h^{-1}\text{Mpc}$. The gray profile represents the original smoothed densities, the orange line is derived from the filtered mock catalogue.

Comparing the two higher order methods with the linear DTFE profile, we may recognise similar features as in the contour plots:

1. the underdense regions of NNFE and Kriging are less noisy (see for example the underdense region at $200h^{-1}\text{Mpc}$)

2. maxima tend to be wider in the higher order schemes.

3. the topology of the unsmoothed reconstruction at higher distances is much smoother.
Figure 5.18– The radially averaged errors. The top figure gives the radial averaged errors for the DTFE method. Both two error definitions are given, $\epsilon_1$ is the full line, dashed lines is $\epsilon_2$ (see equation 5.24). The scales are represented by the colour intensity going from dark $(1h^{-1}\text{Mpc})$ to light $(10 h^{-1}\text{Mpc})$ and the $3$ and $6h^{-1}\text{Mpc}$ scales in between. The lower figure shows the radial average trends according to the second error definition, $\epsilon_2$. The full line is the DTFE method, dashed is the NNFE, and dotted-dashed line in the Natural Lognormal Kriging method.

5.8.3 Intrinsic Smoothing Scale

The last point typically arises in a magnitude selected sample. As one goes out to larger distances the point sample is more and more represented by just the very luminous objects. At the very large distances only the most luminous objects remain visible. If such bright galaxies would be biased in very complicated ways (e.g. higher order biasing), then it would be very difficult to correct for such trends. Here we can only assume that galaxies are tracers of the density field.

To zeroth order we are able to correct for this dilution. That is to say the weighing provides us with zero mean. However the loss of galaxies in the point sample implies an increase of the mean galaxy separation, hence the intrinsic smoothing scale $R_{int}(z)$ also grows with redshift. Therefore the dilution of the point sample causes a loss of the small scale resolution, and $R_{int}(z)$ increases with distance.

In the following we will aim to calculate the errors and argue that the filtered density field is reliable when $R_f > R_{int}(z)$ holds. Unfortunately there is no expression for $R_{int}(z)$ (for example the mean separation is heavily biased towards high density regions), but we may estimate the distance when $R_f > R_{int}(z)$ if there is a systematic change in the errors. Since we would expect these errors to rise more steeply after this transition distance. Other redshift survey will have different transition distances. These could be estimated by rescaling the point sample to the same particle density.

5.9 Quantitative Density analysis

The quantitative comparison we will carry out here is based on the reconstruction of the ideal observed redshift catalogue, i.e. when we would be able to observe all the galaxies (available from the Millennium Mock sample) and a reconstruction from the galaxies that would be observed by a Sloan Survey in the Millennium Universe. Here we will not treat errors induced by peculiar velocity distortions. This will be further investigated in Chapter 7.

In order to quantify the reconstructed density fields, we do a point to point comparison of the density values. Since this is a local comparison, it does not distinguish between a systematic offset or a fluctuating error behaviour. To make such a distinction the average trends of the fields are analysed by comparing the reconstructed fields filtered on a particular scale. We will take a set of four Gaussian filtered fields at scales of; $1.0, 3.0, 6.0$ and $10h^{-1}\text{Mpc}$. These
scales roughly represent the transition from the non-linear scales ($1h^{-1}\text{Mpc}$) to the quasi-linear and linear scales at $10h^{-1}\text{Mpc}$ and larger. First in sect. 5.9.1 we analyse a direct voxel to voxel error estimate. Then in section 5.9.2 we do a correlation analysis of the density values. In the next section 5.10 we will also compare the topology of both fields.

### 5.9.1 Error Analysis

To compare the density fields, the values at each voxel in the full catalogue are compared to ones reconstructed with the mock sample. To quantify the local errors we use the following two error definitions;

\[
\varepsilon_1(r) = \left| f(r) - f(r)_{\text{mock}} \right|, \tag{5.23}
\]

\[
\varepsilon_2(r) = \frac{f(r) - f(r)_{\text{mock}}}{f(r)}. \tag{5.24}
\]

The first error definition is the absolute deviation between the density value reconstructed using the full galaxy catalogue (see figure 5.13) with respect to the mock reconstructed values (see figure 5.14). The second error definition is a relative error. This takes into account that an absolute deviation in underdense regions is relatively a larger error than the same deviation in an overdense region.

In Figure 5.17 the errors are plotted corresponding the density profiles in figure 5.16. The green line is the error from the unfiltered field (the black and the red lines in the upper panels), the purple line is the error profile for the $10h^{-1}\text{Mpc}$ smoothed field. The errors from all three methods are fairly similar. The green line in all cases shows an increasing trend with distance. At a distance of approximately $100h^{-1}\text{Mpc}$ it is of the order of unity. Beyond this distance the errors are characterised by wide peaks that mainly arise by underdense features not represented in the reconstructed mock field.

The errors in the $10h^{-1}\text{Mpc}$ filtered field remains small at an average error level of around 10 percent. In all cases the large scale filtered fields has larger errors at small distances. This is attributed to the mask which is relatively thin (in the z-direction) at short distances ($R < 100h^{-1}\text{Mpc}$). One may notice that the raw reconstructed field at large distance resembles the filtered profile.

We have also plotted the average error behaviour as function of distance in Figure 5.18. These error trends were obtained by averaging all the errors in increasing radial shells. To minimise edge effects we eliminated error data within 15 voxels from the edge. Visual inspection of the error maps showed that this is well sufficient to eliminate errors that derive from edge-effects.

The top panel in Fig. 5.18 displays the radial averaged errors according to the two error definitions for the DTFE reconstruction maps. The dashed lines provide the radial averaged errors for the second error definition. Especially at small scales the errors by the second definition are systematically higher. This indicates that the errors at those scales are dominated by the underdense regions. At larger scales the difference vanishes and both error definition provide almost equivalent error trends, hence at larger scales errors are more equally distributed over a range of density values.

At distances closer than $100h^{-1}\text{Mpc}$ the average error trend is reversed and errors decrease with distance. This trend may originate from the relatively small volume that is probed at close-by distances, and is most prominent on the largest filter scales. Here too less independent wavevectors are available to fully reconstruct the density field. This directly influences the reconstruction, but also indirectly via the weight factors. These weights are derived from a
fit to the selection function and at close-by distances this fit is heavily influenced by the presence or absence of large superstructures. Differences in the fit to the selection function will cause systematic errors via the local density estimates. For example is the weight factors are too low then reconstructed field could exhibit a systematic offset above the mean value. Since we may only observe one Universe such systematics are probably best dealt with by using only volume limited samples at close-by distances \((R < 150h^{-1}\text{Mpc})\).

The bottom panel of Fig. 5.18 we show the error \(\langle \varepsilon_2 \rangle\) trends for the four scales and all three methods. For all three methods the errors at the smallest scales \((1h^{-1}\text{Mpc})\) are high and never better than 50%. Fortunately the errors decrease quite rapidly with filter scale. At large scales \((R_f > 6h^{-1}\text{Mpc})\) the density field can be reconstructed with quite high accuracy. Throughout the whole volume (discarding the first 100\(h^{-1}\text{Mpc}\)) the relative errors will be no more than 10%. In upcoming studies we will investigate until what distance the largest scale can be accurately reconstructed. A reliable reconstruction at high redshift could provide a mean to probe the cosmological growth factor. Overall strikingly little differences are found between the methods. None of the methods does perform exceptionally better than any of the other. Looking at the details we may only observe two differences. The first is that the Kriging errors at small scales are relatively high compared to the other two. Secondly, at larger filter scales there is a slight, but systematic preference for the DTFE method. Hence the most simple and straightforward reconstruction method seems to work best.

### 5.9.2 Density field Correlation

Another way to look at the performance of the density reconstruction is by means of a correlation analysis. Such an analysis may reveal whether errors are dependent on density. Also systematic errors like deviations from linearity or offsets can be uncovered with these diagrams.

In figure 5.19 we plot for the DTFE reconstructed fields the correlation diagrams at filter scales of \(R_f = 0, 1, 6, 10h^{-1}\text{Mpc}\). The volume that were compared are given in section (9.1.2). Each of the contour diagrams was made by counting the number of voxels in the logarithmic density - density plane. The plane was sampled using 200 bins on each side, the contour levels where chosen to start from one pixel per bin (black) until a maximum (white). The black line provides the perfect one-to-one relationship between the (full) density and the reconstructed density.

In general the correlation is characterised by a linear slope. This slope is perfectly centred on the one-to-one relation at large scales and slightly offset at the smallest scales. The latter is a reflection of the systematic offset at nearby volumes. As already described in the previous section 5.9.1 it is a reflection of the weighting scheme.

Naturally the scatter on the linear correlation is highest at the smallest scales \((1h^{-1}\text{Mpc}\) and \(3h^{-1}\text{Mpc}\) (Figure 5.20). Also at these scales the variance in the correlation cloud is highest at low densities. This reaffirms our previous conclusion that at small scales the uncertainty in low density regions is larger than in high density regions. Even though the variance of the point cloud increases the overall mean trend remains very much a linear relation.

At large scales the scatter is more or less uniform and centred perfectly on the one-to-one relationship. The slight deviation at lower density values might indicate a systematic bias towards more underdense values. On the other hand the low and high density extremes are governed by only a few rare objects. Alternatively it could therefore indicate a finite volume effect, i.e. cosmic variance. Since this deviation is not much larger than the intrinsic spread we do not consider it as a serious problem.

In figure 5.20 we plot the density correlation for the three methods at a filter scale of \(R_f = 3.0h^{-1}\text{Mpc}\). Each method shows a correlation that is over two orders of magnitude in density
Figure 5.19— Correlation diagrams between the density reconstruction from the full galaxy catalogue ($\delta_{\text{full}} + 1$) and the density reconstruction from mock catalogues ($\delta_{\text{mock}} + 1$) at several scales. The colour indicates the number of voxels of the density fields that occupy a certain area in the correlation diagram. The contour levels are distributed using a powerlaw model with index $\approx 2.5$. The four panels show the correlation at four different filters scales. The top-left figure is for the unsmoothed reconstructions ($R_f = 0$), top-right for $1.0h^{-1}$Mpc filtered fields, and the bottom-left and the bottom-right figures are respectively for a filter radius of $6.0$ and $10h^{-1}$Mpc fields.
Figure 5.20— Similar as in figure 5.19 we plot the density correlation between the full catalogue and densities of the mock catalogue. For the three different methods, DTFE (left), NNFE (middle) and Natural Lognormal Kriging (right). All fields were smoothed at a filter scale of $3h^{-1}\mathrm{Mpc}$ and the same volume was used for the comparison.

Figure 5.21— The correlation between unsmoothed full density and the unsmoothed mock density field. The black contours shows the cross-correlation between the $3.0h^{-1}\mathrm{Mpc}$ filtered field of the full catalogue and the unsmoothed mock density field.

is almost perfectly a one-to-one relationship. In other words consistent with being unbiased. Furthermore we find no large differences between the methods. Also the correlation analysis and the error analysis are in agreement and show that the reconstruction method does not reduce nor increase the errors in the density field.

On this basis we suggest that the errors in the reconstructed fields are fully dominated by the errors of the (initial) density estimator (an alternative explanation will be considered and rejected in section 5.9.4). Since this is a fundamental limit of the observed point distribution, a substantial improvement would either require a deeper redshift surveys, or a improved (initial) density estimator.

5.9.3 Intrinsic Smoothing & Nonlinearities

From the reconstructed contours we got the impression that the reconstructed field resembles a filtered field. To quantify this effect we plot in Figure 5.21 the correlation between the unsmoothed field from both the full sample and the mock sample. The correlation between these is indicated by the coloured contours. It may be obvious that although there is some degree of correlation it is not tight, the scatter is large at almost all densities. The black contours
Figure 5.22— The radially averaged error for the DTFE reconstruction method for the mock (or magnitude limited sample) and the volume limited sample (dashed). The top figure gives the errors according to the first error definition $\epsilon_1$, the lower figure is for the second error definition $\epsilon_2$. In each panel we give the error trends for three different filter scales, 1.0, 3.0 and 10 $h^{-1}$Mpc.

Present the correlation between the full catalogue filtered at $3.0 h^{-1}$Mpc and the unfiltered mock density field. This correlation is much tighter, and underlines the overall impression from figures 5.13 to 5.16 that the lack of galaxies implies a smoothing of the underlying field.

In addition to the rather moderate correlation we find that the relation is not linear, as is manifest in from curved mean trend. Reconstructed overdense values in the mock sample are systematically mapped to higher values with respect to $3h^{-1}$Mpc filtered field. The average to moderately low density values appear to mapped to lower values in the mock sample, while a reverse trends may be seen at the very low densities. In other words the intrinsic smoothing does not only depend on distance, but also on density. This means that the intrinsic smoothing $R_{\text{int}}(z)$ is larger at low densities. Another way of seeing this is that most of the information is contained in objects with a high number of galaxies.

Applying a filter on the raw reconstructed density field will smooth out the high density values, this mass gets redistributed into the lower density regime. We argue here that this redistribution can only be done correctly, when the nonlinear features in the raw field were reconstructed at the correct position and with the correct amplitude. After filtering and redistribution we obtain the tight correlation visible in Figure 5.19 and Figure 5.20. This implies that information from high density objects or nonlinear sites is crucial to reconstruct the large scale density field. At small filter scales much less information from high density object is able to reach the low density regime, producing the increasing error trend (Figure 5.18 and Figure 5.20) towards lower densities.

5.9.4 Magnitude limited versus Volume limited

Considering the slow radial increase of the errors with distance, it is worthwhile to ask the question how the errors compare to those generated from a volume limited samples. In a magnitude limited we have to correct for the steady dilution of the point sample. Such corrections may introduce new errors and/or systematic differences. Some of these were already found at close distances. In order to test this we constructed a volume limited sample from the Delucia Millennium mock catalogue, between a redshift of 0.02 and 0.1.

An example for each of the three density reconstructions from this sample was shown in Figure 5.14. In Figure 5.15 we have shown the three mock samples using the NNFE reconstruction (top: full sample, middle: magnitude limited sample, bottom: volume limited sample). The loss of small scale details in both mock samples is obvious. Fine filamentary features that separate small voids are either merged into larger overdense complexes or lost into observationally enlarged voids.
Figure 5.23— The correlation between the density of the full catalogue and densities of the magnitude limited mock catalogue is shown in the left figure. Right shows the full catalogue compared to the density of the volume limited catalogue. All fields were smoothed at a filter scale of $3h^{-1}\text{Mpc}$.

Figure 5.24— We compare the volume limited errors of the three different methods at the four different scales in consideration (from black to orange; 1.0, 3.0, 6.0, and 10 $h^{-1}\text{Mpc}$). The radial averaged errors ($\epsilon_2$) by the DTFE method are indicated with the solid lines, NNFE errors with the dashed lines, and the dot-dashed line shows the errors for the Natural Lognormal Kriging reconstruction.

For the volume limited samples we have repeated the error and correlation analysis. The error analysis can be found in Figure 5.22. It shows both the absolute deviation and the relative deviation for the DTFE reconstruction. Volume limited samples are statistically uniform, so that we expect an error trend of a constant error value as a function of distance. This is indeed what we find. The trends for the magnitude limited samples were already discussed in the previous section. Comparing the volume limited with the magnitude limited trends we find the magnitude limited errors are consistently smaller than for the volume limited sample.

Figure 5.23 shows the correlation diagrams for the magnitude (left) and volume limited samples (right). Overall they seem to be quite similar. However there important differences in the details, one concerns the maximum (white areas). For the magnitude limited this has a much more extended concentration. Providing on average smaller errors for the magnitude limited sample. At approximately the last measured radial shell, the point sample of the magnitude and volume limited samples are the same. This implies that the errors at furthest measured shell (around a distance of $300h^{-1}\text{Mpc}$) in the volume limited sample should be approximately equal to magnitude limited sample. Figure 5.22 provides a visual confirmation of this picture.

We compare the performance of the three methods on the volume limited sample (see
Figure 5.25— The top panel shows the masked Millennium density field. Superimposed are the WVF-watershed lines derived from the millennium mock sample. The bottom figure shows the segmentation for the full sample (red lines) as well as the mock sample (black lines). The topological errors are indicated by the various colours: orange for false mergers and red for false splits (see text for explanation).

The radially averaged errors ($\epsilon_2$) are plotted for the four filter scales and the three methods. It confirms the same conclusion found in the previous section; we hardly find any differences between the methods. This indicates that the errors are caused by the estimates in the density values. Note that this excludes a possible alternative explanation that we did not mention previously. In principle the weights in magnitude limited samples are only correct for the DTFE method. The weights could have increased the errors for the higher order methods. However given the fact that the minor differences between the three methods in Figure 5.24 are very similar to those in Figure 5.18 Concluding the simplest and most straightforward method (DTFE) performs either as well as, or even better than the two higher order methods.
Figure 5.26— A depiction of a false-split. The four gray circles represent the patches from the reconstructed segmentation, the black circles would refer to a patch in the original segmentation. The gray shaded areas belong to the areas of the three reconstructed patches ($f_R$) that have an intersection with the black circle. The dashed areas represents intersections between the original and the reconstructed segmentation ($f_{R/O}$).

5.10 Topological Comparison

From figure 5.14 we noticed that the contours of the higher order methods were much smoother. This did not translate into an improvement of the errors. As an alternative we will now examine how well the the topology of the field was reconstructed. Even when two fields have similar errors according to Eq. 5.11, the topology can still be different. This provides another independent measure of the quality of the reconstructed density field.

We will measure the topology according to the minima and their surrounding void-patches. Again the comparison is based on the density field reconstructed with the full sample and the (magnitude limited) mock-sample. The topological errors are quantified according to the watershed segmentation of both fields.

Figure 5.25 shows the watershed segmentation generated by the mock catalogue (the black lines). It is plotted on top of the full sample density field. The bottom figure shows the segmentation according to the full sample (red) and the segmentation of the mock sample is plotted in black. The lower figure shows two very obvious cases where the differences between the two segmentations is large. In the two large underdense regions at respectively $X=180$, $Y=180$ and $X=450$ and $Y=225$, the smaller bridges between the sub-voids have vanished. Such effects are to be expected, since at larger distances the only information that will be retained is the skeleton of most overdense filaments and walls traced by the brightest objects. The smaller filigree of filaments inside low density regions is lost. We will now try to quantify this loss of structural information and see whether there are any differences between the three methods.

5.10.1 Quantification of Topological Errors

Computational Definition of Topological Errors

The errors in the segmentation are found by comparing patches of the two segmentations, see also Platen et al. (2007). Such errors can be classified to first order into false-splits and false-mergers. A false-split is defined when a patch from the ideal density field is splitted into two or more voidpatches. Vice verse a false-merger is then defined when two or more patches from the original segmentation are merged into one object in the mock reconstruction, see Figure 5.26 for a visual representation of a false-split.

Computationally we define a false-split in the following manner; for each area in the original field (black circle in Figure 5.26) find the patches from the reconstruction that overlaps with it. In Figure 5.26 these would be the three gray shaded circles belonging the void1, 2 and 3. For each patch calculate its volume, $f_R$, and of the intersecting volume, $f_{R/O}$. Then
Figure 5.27— The (DTFE) radial profiles are plotted for the percentage of correctly recovered void-patches (full) as well as the percentage of topological errors (dashed). The various colours represent the different scales, resp. $1.0 \, h^{-1}\text{Mpc}$ (dark blue), $3h^{-1}\text{Mpc}$, $6h^{-1}\text{Mpc}$ and $10h^{-1}\text{Mpc}$ (lightest blue).

according to the ratio, $\frac{f_{\text{no}}}{f_{\text{g}}}$, we determine whether an overlap is significant or not. Here we assume that when the ratio is larger than 0.6 then an overlap is a significant. For example, this excludes the case of void3 in Figure 5.26 as it represents a slight shift of the boundary (not all morphological changes need to be of a topological nature). Then a \textit{false-split} is defined when there are at least two or more patches with a ratio larger than 0.6. In the example of figure 5.26 void1 and void2 both have ratios higher than 0.6. Hence there are at least two patches with significant overlap and this situation is classified as a false-split. A false-merger is computed in the same way, but then the roles of the two segmentation are reversed. Since by symmetry a false-merger can be considered a false-split in the original segmentation.

Having defined the errors we can now define the correctly identified patches. A correct patch is computationally defined when there is only one object with an overlap fraction of more than 60%. This has to hold for both the patches in the reconstructed field as well as the corresponding patch in the original density field. This prevents a patch from being either a false-split or a false-merger.

\textbf{Topological Error Comparison}

In the lower panel of Figure 5.25 we plot the spatial distribution of the topological errors. Both the original segmentation (black lines) and segmentation from the mock sample (red lines) are plotted. The false mergers are indicated by the orange patches, the red patches are false splits, and the correct patches are shown white. From the plot it is apparent that false mergers are more abundant than splits. If we compare the two panels we may observe that the segmentation lines preferentially disappear inside void regions. This means that minima inside large scale void regions, are merged into one object. This preference can be easily understood if one considers that walls inside larger void regions are more underdense. Only a few galaxies have to disappear to be able to merge two voids. On the other hand the coherence of the reconstructed watershed lines with the full density field, remains strong throughout the volume.

The number of false splits and false mergers as function of distance can be used as an measure for how well the density field is reconstructed. Similar to the errors in the previous section we will count in radial shells the number of voidpatches, the number of error-patches ($\text{false splits} + \text{false mergers}$) and the number of correctly identified patches. This allows us the calculate the percentage of correctly identified patches and the percentage of misidentified
Figure 5.28— The $3 h^{-1}$Mpc (left) voidpatch recovery for DTFE (blue), NNFE (orange) and Kriging (gray). The full line represents the percentage of correct identifications, the dashed lines is the number of incorrect identifications. The lower panels shows the performance of the three interpolation methods at filter scale of $10 h^{-1}$Mpc.

regions with respect to all the voidpatches as function of distance.

In Figure 5.27 we plot the percentage of correct and incorrect identified patches for the DTFE reconstruction at the filter scales of 1, 3, 6, 10 $h^{-1}$Mpc. The number of correct identifications at the small scales is always extremely low and drops off quite sharply. Even the topology at $3 h^{-1}$Mpc can only be reliably reconstructed within a distance of only $200 h^{-1}$Mpc. For the $6 h^{-1}$Mpc and $10 h^{-1}$Mpc the performance is slightly better, but the drop down the 50% line at 300 and $400 h^{-1}$Mpc. This is a reflection of the fact the errors are most pronounced in low density environments.

In Figure 5.28 (right panel) we plot the percentage of both the correct number of recovered voidpatches as well as the percentage of topological errors (false splits + false mergers) for the DTFE, NNFE and Kriging reconstruction. The two plots show the radial recovery for two scales $3 h^{-1}$Mpc and $10 h^{-1}$Mpc. Although the underdense regions are smoother for the higher order methods, this has not resulted in a higher number of correct identified voidpatches. Overall there is again a slight preference for the DTFE reconstruction.

In general we notice that the topological errors increase with distance and consequently the number of correctly identified patches decreases sharply. The point when both trends are equal provides an estimate for the maximum distance out to which we still may trust the reconstructed field. Using Figure 5.27 we obtain the following estimates for the maximum distance $R_{\text{max}}$:

\[
R_{\text{max}} \approx 100h^{-1}\text{Mpc at } R_f = 1h^{-1}\text{Mpc}
\]
\[
R_{\text{max}} \approx 200h^{-1}\text{Mpc at } R_f = 3h^{-1}\text{Mpc}
\]
\[
R_{\text{max}} \approx 300h^{-1}\text{Mpc at } R_f = 6h^{-1}\text{Mpc},
\]
\[
R_{\text{max}} \approx 400h^{-1}\text{Mpc at } R_f = 10h^{-1}\text{Mpc}
\]

Within these maximum distances also the errors (section 5.9.1) are well below unity.
5.11 SDSS-DR6 density reconstruction

In Figures 5.29 & 5.30 the density reconstruction is shown for the 7th data release of SDSS. For comparison we plot in figure 5.29 the galaxies from which the density field was derived. The density reconstruction according to the NNFE method and the Natural Lognormal Kriging method are shown in Fig. 5.30. Besides some minor differences, overall one is able to find the same prominent structures in all three density maps. Again both higher order methods provide very similar and smooth density fields. In the next Chapter 6 we will extensively discuss the cosmography of these density fields.

5.12 Improvements

5.12.1 Mass Conservation & Density Estimation

The tested density reconstruction methods all started out with the same density estimate (see section 5.3). This estimate is the correct choice for the DTFE, as it guarantees mass conserva-
For NNFE there is no such direct connection with the density estimator, and one is not guaranteed to conserve mass. One could attempt to normalise the NNFE kernel, however this kernel is much more complicated than the linear DTFE kernel. This approach would only be viable if the mass of the local NNFE kernels can be computed relatively fast.

The simple Kriging method is by construction mass conservative, since the mean of the field is constrained to be zero. However this is only valid if the method is implemented globally. With the localised version we may have given up this mass conservative estimation. And the consequence of localisation on the mass conservation should be investigated more rigorously (below we suggest a hierarchical and mass conservative implementation). In general the Kriging method is more flexible than the other two methods with the choice of the density estimator. Therefore it may be worthwhile to explore whether other density estimators provide better reconstructions. These could for instance give an improved handle on the shot noise in the density estimate.

---

**Figure 5.30**—Same as figure 5.29, but now for the SDSS field reconstructed with NNFE (top) and Kriging (bottom). The fields were derived from the same point set.
5.12.2 Grid Improvements

The regular grid (cubic) played an important role in the density field. Not only is it the most convenient way to store the data, but also for the post-processing one often requires that the data is on a grid (e.g. filtering operations). Despite their advantage, regular grids also have huge disadvantages. They have great difficulty representing hierarchical data, since any grid imposes a resolution scale. Another large problem of regular grids is their poor handling of anisotropic, or nontrivial directional dependent data. Considering that the Cosmic Web is both highly hierarchical and very anisotropic one does need to worry about such grid based representations.

Here two immediate improvements can be made over the regular grid that we have used. The first improvement would be the use of a polar grid instead of a regular cubic grid. Especially given the observational imposed geometry, i.e. angular sky coverage and the radial dependence of the selection function. Such a spherical grid would immediately lead to a better representation of the observation mask, redshift distortions of clusters (fingers of god), the void stretching and the selection function. Here we could naturally increase the radial binsize with distance. In a magnitude limited sample it naturally adapts the binsize according to the radial decrease of the number of observed galaxies. Here we would obtain higher resolution at close distances and a coarser resolution at further distances.

Another point is the integration of the density field within a voxel element. As pointed out in the Appendix 5.F in case of a spatially irregular sample the natural Nyquist frequency or its real space counterpart is the Voronoi cell. To avoid aliasing in the sampling of the density field on a grid we have to oversample grid cells. An exact alternative does exist; the piecewise density integration of the tetrahedra that are embedded within the voxel. For DTFE this can be done exact and relatively fast. The computational aspects are not trivial, since one needs to take into account the intersections between the tetrahedra and the voxels.

5.12.3 Kriging Improvements

For the Kriging method we can see three areas for improvement; the neighbourhood, the data-model and the matrix inversion. The implemented running neighbourhood will still try to interpolate the discontinuities within the local neighbourhood. An alternative are predictor-corrector schemes that decide locally which data points to take into account or not. This avoids both the smoothing (SPH) and the oscillations (Kriging) in and around discontinuities. In hydrodynamics such schemes are called essentially non oscillatory ENO-schemes, see Harten & Osher (1987) and Abgrall (1994).

Related to the choice of the neighbourhood is the problem of the large matrix inversion. Localising the set of data points circumvents the issue of computational overload. However by doing so we may also neglect some of the large scale correlations. It may introduce biases in the reconstruction. There has been some attention to reconstruct fields at various resolution scales using resolution tree structures (see Magnussen et al. 2007, and references therein). Such multi-resolution approaches often uses a partition of unity, the requirement that high resolution patches sum up to one at the lowest resolution. This guarantees a proper normalisation within the whole range of scales. Combining these multi-resolution trees with the above predictor-corrector scheme may provide a criterion for when to open or not to open a leaf node. For example, one may determine not to refine the resolution when a node ventures into a particularly strong non-linear object. Such an implementation may solve in a natural manner the issues addressed in this study but goes beyond the scope of this work.

In order to partially deal with the nonlinearities we transformed the data into the logarithm of the data values. As became clear in this work this provided a much better behaved density interpolation. It may be worthwhile to explore whether there are other transforma-
tions that are even superior. One possibility is a transformation using the Hermite Polynomials. These can transform any density distribution into a new density distribution with a Gaussian One Point Probability Distribution. We already briefly mentioned this form of Kriging with the name disjunctive Kriging (sect. 6.4). In Chapter 7 we will discuss that the density distribution on small scales deviates from a lognormal distribution. This suggests that a combination of a logarithmic transformation followed by a further correction using the Hermite Polynomials would be a logical next extension of the Natural Lognormal Kriging.

5.13 Summary and Future Work

For many applications the reconstruction of the density field is a vital first step in the analysis of the Large Scale Structure. In this study we address two of the idiosyncratic properties of density fields in a cosmological context. The first is the fact that we have to attempt to infer the density fields from a spatial distribution of discrete objects, usually galaxies in an observational setting or particles in an N-body simulation. The second issue concerns the spatial properties of the density distribution. Characteristically hierarchical, it involves a large range of spatial scales, densities and masses. In addition, it is marked by highly anisotropic filamentary and sheetlike geometries, integrated in a pervasive weblike network.

In order to be able to study and analyse in an unbiased and objective fashion the various properties mentioned above we here evaluate three particularly suited reconstruction procedures. One particular field of application is the analysis of large galaxy redshift surveys, such as the SDSS, with the intention of understanding its statistical properties and to isolate and identify features such as voids, filaments and clusters.

Here we divide the reconstruction process up into two parts. the first part is the estimate of the density for each galaxy. The second step involves the interpolation of the data to obtain a density field. Preferably, both steps should be coupled, guaranteeing the conservation of mass of the reconstructed field.

We have used three different interpolation methods; the Delaunay Tessellation Field Estimator (DTFE), Natural Neighbour Field Estimator (NNFE) and a local implementation of Kriging, the Natural Lognormal Kriging. DTFE, NNFE and Kriging are well known techniques for interpolating scattered data. In this study we extended Kriging to a formalism restricting its correlation to a neighbourhood of natural neighbours, resulting in the Natural Lognormal Kriging method.

All three methods involve a local neighbourhood of natural neighbours, i.e. tetrahedral natural neighbourhood based on the Delaunay triangulation. This allows the local methods to adapt automatically to the underlying point distribution. One immediate benefit is that it significantly speeds up the reconstruction method. As may be understood from the tests and results in this study, such an approach is able to deal with a variety of artifacts indigenous to discretely sampled density fields. Cosmological density fields are not always well behaved. They contain numerous non-linear objects and structural singularities, involving locations where is not properly differentiable.

One important issue which we call peak correlated overshoots. We have shown that numerical artifacts occur when such nonlinearities are globally interpolated using higher order methods. In a simple one dimensional example we visualise this effect when a nonlinearity is introduced into otherwise well-behaved data. The artifacts appear as extreme oscillations around this high density peaks related to well known Gibbs phenomenon. As we demonstrated here, by restricting the interpolation to the local neighbourhood of the nonlinear data, we end up with a more robust solution for the above problem. The Gibbs oscillations are strongly reduced and produce a more well behaved interpolation method.

These three density reconstruction methods were analysed and compared using mock
SDSS redshift samples. For each method the (relative) errors trends, biases and topology were investigated. We showed that for a range of scales and corresponding distances the density field can be reliably reconstructed. Density fields filtered over large scales retain a better coherence with the underlying field, implicitly also out to much longer distances. Especially in the larger volume of magnitude limited samples the errors increase only very slowly.

We have indicated the importance of an accurate reconstruction in the non-linear regime. First and foremost, because of the above mentioned oscillatory errors. Secondly, because the accurate reconstruction of nonlinear sites allows the unbiased inference of the density field at larger scales. This despite the fact that these sites may seem to be relatively ‘small’ objects compared to the larger scale density field from which they are ‘decoupled’. This is not the case, since rare high density peaks occur at the seldom interference of both small and large wavelengths (Bond & Myers 1996). These sites are the tracers of the large scale field. We conclude that nonlinearities are not only of importance to reconstruction of small scales, but also vital for an accurate reconstruction of the large scale density field.

On theoretical grounds one might have expected that higher order methods should provide a better reconstructed density field. Cosmetically the two higher order methods do produce more visual appealing reconstructions. However the three methods have very similar errors. Both the errors in the local density value as well as the errors measured via the topology show almost the same trends. We attribute this to the uncertainty in the initial estimated density for each galaxy. A higher quality reconstruction thus requires either a better density estimate or deeper observational data.

We found that the DTFE method performs marginally better than the other two methods. A possible explanation for this somewhat counter-intuitive result is that higher order methods respond in higher order fashion to (intrinsic) errors. In such a situation the best reconstructed filtered field is achieved when the unfiltered field was reconstructed with the smallest linear errors. An additional factor might be that DTFE is the least global. It uses information from four vertices, which is a lot less than the average of 17 and 57 vertices in respectively NNFE and Natural Lognormal Kriging. Estimated density errors are therefore present in typically much smaller volumes. As a consequence they do filter out more quickly than they do with the other two methods.

The removal of only a couple of void galaxies leads to the spurious merging of observed voids. The topological reconstruction at small scales is therefore relatively difficult. For example at maximum only half or less of the $1h^{-1}$Mpc defined voids could be recovered correctly. A significant improvement is achieved by going to larger filter scales, $3h^{-1}$Mpc or larger. An implication is that a proper analysis of the small scale voids and the galaxies therein can only be achieved within the first $100h^{-1}$Mpc. Within this volume the small scale $1h^{-1}$Mpc field is still reconstructed within tolerable errors. The environment defined by these small scales seem to be most relevant for galaxy properties (Park et al. 2007). When one is interested in the relation between galaxy properties and local (nonlinear) small scale environment one is constrained to use only this nearby volume. Otherwise environmental confusion, the process of filament and wall galaxies to preferably end up in voids, will erase the possible delicate relation between galaxies and small-scale environment.

In this work we based ourselves on spectroscopic galaxy redshift samples, in which the radial errors are dominated by redshift distortions. However many of the larger upcoming surveys will be using photometric redshifts with larger expected radial uncertainty than spectroscopic samples. We suggest that the DTFE method could also be applied on such samples. Even though small features are erased we suspect that the density field at larger scales ($R_f > 10h^{-1}$Mpc) can still be reconstructed relatively accurate.
Acknowledgements

EP likes to thank Alex Szalay for his hospitality and financial support at JHU. Also we are thankful for the help from Gerard Lemson and Sebastien Heinis with the Millennium Database queries. We would like to thank Istevan Szapudi and Mark Neyrinck for their helpful discussion.

The Millennium Simulation databases used in this paper and the web application providing online access to them were constructed as part of the activities of the German Astrophysical Virtual Observatory (GAVO).

The SDSS is managed by the Astrophysical Research Consortium (ARC) for the Participating Institutions. The Participating Institutions are The University of Chicago, Fermi-lab, the Institute for Advanced Study, the Japan Participation Group, The Johns Hopkins University, Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, University of Pittsburgh, Princeton University, the United States Naval Observatory, and the University of Washington.

5.A SDSS Coordinate System

The transformation of the DR6 NGP sample and the complete DR7 sample, from the ra-dec-redshift coordinates ($\alpha, \delta, z$) to a grid based ($X, Y, Z$) coordinates (normalised to a box size of $600h^{-1}$Mpc) were done according to;

$$
X = \frac{R(z)}{600.0} \cos(\delta) \cos(\alpha - 90) \\
Y = \frac{R(z)}{600.0} \cos(\delta) \sin(\alpha - 90) \\
Z = \frac{R(z)}{600.0} \sin(\delta), 
$$

(5.25)

Where $R(z)$ is the comoving distance and the offsets is to get the observer cone in the middle of the XZ-plane. A further rotation to get the centre of the DR6 NGP down in the direction of the Y-axis, we adopted these two rotation angles; one rotation around the X-axis $\eta = 47$ and one further around the Y-axis $\nu = 3.6$.

$$
X' = X \\
Y' = \cos(\eta)Y + \sin(\eta)Z \\
Z' = \cos(\eta)Z - \sin(\eta)Y
$$

(5.26)

$$
X = \cos(\nu)X' - \sin(\nu)Z' + 0.5 \\
Y = Y' \\
Z = \sin(\nu)X' + \cos(\nu)Z' + 0.5
$$

(5.27)

5.B DTFE implementation

This is an outline of the step in the DTFE procedure for reconstructing density field from a galaxy sample.
• **Point sample** Given that the point sample is supposed to represent an unbiased reflection of the underlying density field, it needs to be a general Poisson process of the (supposed) underlying density field.

• **Boundary Conditions** The boundary conditions for a galaxy sample; *Vacuum boundary conditions*: outside the sample volume there are no points. We have used in all cases vacuum boundary conditions.

• **Delaunay Tessellation** Construct the Delaunay tessellation from the point sample using the Computational Geometry Algorithms Library (CGAL) library.

• **Field values point sample**

  The estimate of the density at each sample point is the normalised inverse of the volume of its contiguous Voronoi cell $W_i$ of each point $i$. The contiguous Voronoi cell of a point $i$ is the union of all Delaunay tetrahedra of which point $i$ forms one of the four vertices. The *Systematic non-uniform sampling process*: sampling density according to specified selection process. The non-uniform sampling process is quantified by an a priori known selection function $\psi(r)$. This situation is typical for our galaxy surveys, $\psi(r)$ may encapsulate differences in sampling density $\psi(\alpha, \delta)$ as function of sky position $(\alpha, \delta)$, as well as the radial redshift selection function $\psi(r)$ for magnitude- or flux-limited surveys. For $D$-dimensional space the density estimate is,

  $$\rho_r = (1 + D) \frac{w_i}{\psi(r) V(W_i)}.$$

• **Field Gradient**

  Calculation of the field gradient estimate $\nabla f_m$ in each $D$-dimensional Delaunay simplex $m$ ($D = 3$: tetrahedron; $D = 2$: triangle) by solving the set of linear equations for the field values at the positions of the $(D + 1)$ tetrahedron vertices,

  $$\nabla f_m \iff \begin{bmatrix} f_0 & f_1 & f_2 & f_3 \\ r_0 & r_1 & r_2 & r_3 \end{bmatrix}$$

  Evidently, linear interpolation for a field $f$ is only meaningful when the field does not fluctuate strongly.

• **Interpolation.**

  The final step of the DTFE procedure is the field interpolation. The processing and post-processing steps involve numerous interpolation calculations, for each of the involved locations $\mathbf{r}$. Given a location $\mathbf{r}$, the Delaunay tetrahedron $m$ in which it is embedded is determined. On the basis of the field gradient $\nabla f_m$ the field value is computed by (linear) interpolation,

  $$\tilde{f}(\mathbf{r}) = f(r) + \nabla f_m (\mathbf{r} - r_i).$$

  In principle, higher-order interpolation procedures are also possible. Two relevant procedures are:
- Spline Interpolation
- Natural Neighbour Interpolation Watson (1992); Braun & Sambridge (1995); Sukumar et al. (1998); Okabe et al. (1992)

- **Processing.**
  Though basically of the same character for practical purposes we make a distinction between straightforward processing steps concerning the production of images and simple smoothing filtering operations on the one hand, and more complex post-processing on the other hand. The latter are treated in the next item. Basic to the processing steps is the determination of field values following the interpolation procedure(s) outlined above. Straightforward “first line” field operations are “Image reconstruction” and, subsequently, “Smoothing/Filtering”. Image reconstruction. For a set of image points, usually grid points, determine the image value: formally the average field value within the corresponding gridcell here we use Monte Carlo approach

*Smoothing and Filtering:*

- Linear filtering of the field \( \hat{f} \): convolution of the field \( \hat{f} \) with a filter function \( W_f(\mathbf{r}, y) \),
  usually user-specified,
  \[
  f_s(\mathbf{r}) = \int \hat{f}(\mathbf{r}') W_f(\mathbf{r}', y) d\mathbf{r}'
  \]  \hspace{1cm} (5.31)

- **Post-processing.**
  The real potential of DTFE fields may be found in sophisticated applications, tuned towards uncovering characteristics of the reconstructed fields. An important aspect of this involves the analysis of structures in the density field. This can be finding voids, identifying cosmic structures or retrieving the One-Point Probability Density Function.

## 5.C Radial Basis Interpolation

A very promising method for interpolation is that of the Radial Basis interpolation. It is one of the most the leading methods for interpolating irregularly spaced data. The principle of RBF is the approximation of an unknown function \( \hat{f} \) by a linear combination of basis functions \( \phi(\mathbf{r}_i, \mathbf{r}_j) \). When these basis functions depend only on distance, then they are referred to as Radial Basis Function;

\[
\hat{f}(\mathbf{r}) = \sum_{i=1}^{N} \lambda_i \phi(\mathbf{r}_{ij}).
\]  \hspace{1cm} (5.32)

The values \( \lambda_i \) are determined from the following linear system of known values,

\[
f_i = \lambda_i A_{ij}.
\]  \hspace{1cm} (5.33)

The entries of this matrix are \( A_{ij} = \phi(\mathbf{r}_i, \mathbf{r}_j) \). When the function \( \phi \) is positive definite then the matrix \( A_{ij} \) is not singular and thus solvable. Some RBF’s do not have this property, then adding a set of polynomials may create a solvable the linear system. A nice property of RBF interpolation is, that it approximates the unknown function in a well-behaved and smooth manner.

There are many choices of radial basis function e.g. (inverse) multi-quadratics, (various) splines or Gaussian, each of with their own merits and culprits. For example the thin-plate
spline function $r^2 \log r$ (2D) and $r^3$ (3D) minimises the total curvature. The tests of Franke (1982) showed that the multi-quadratic RBF provided the best overall performance. In general there is no single prescription for choosing a specific kind of RBF. Testing the various function and scales seems to be the best approach, see Press (2007) for further details and discussion.

Radial basis function interpolation has been used extensively in various applications. One important practical consideration is how to get fast solution of the inversion of the matrix $A_{ij}$. Since this also applies to our situation here, and we like to mention two of those approaches. The matrix $A_{ij}$ is not sparse, because none of the distance terms vanish beyond a certain distance. Fortunately, there are RBF’s (Wendland 1995) that have compact support, providing sparsity of the matrix. The inversion of such matrices far more easily, providing a significant speed up. Another approach is based on the fast multipole method (Carr et al. 2001). Similar to gravitational force calculations long distance terms can be grouped together using a data tree. A suitable hierarchical tree may thus average the data contributions of the RBF’s at large distances. The essential idea in each case is: localise the problem as much as possible, while keep the global relationship of the data intact.

5.D Derivation of the Kriging equation

The Kriging equations can be derived by doing a spatial least squares estimate. Thus minimising the expected squared differences of equation 5.11. The latter can be rewritten as:

$$E(\hat{f}(\mathbf{r})^2) - 2E(\hat{f}(\mathbf{r})f) + E(f^2)$$

(5.34)

using the estimate of eq. 5.10 gives

$$E\left(\sum_i \sum_j \lambda_i \lambda_j f(\mathbf{r}_i)f(\mathbf{r}_j)\right) - 2E\left(\sum_i \lambda_i f(\mathbf{r}_i)f(\mathbf{r})\right) + c(0)$$

(5.35)

from which it follows that

$$\sum_i \sum_j \lambda_i \lambda_j E(f(\mathbf{r}_i)f(\mathbf{r}_j)) - 2 \sum_i \lambda_i E(f(\mathbf{r}_i)f(\mathbf{r})) + c(0)$$

(5.36)

leading to

$$\sum_i \sum_j \lambda_i \lambda_j C(\mathbf{r}_i, \mathbf{r}_j) - 2 \sum_i \lambda_i c(\mathbf{r}_i, \mathbf{r}) + c(0)$$

(5.37)

where $c()$ Then the weights $\lambda_i$ that provide the smallest error are those that have equate the derivative w.r.s.t. the weights to zero:

$$\left.\frac{dE(\hat{f}(\mathbf{r})^2) - f(\mathbf{r})^2}{d\lambda_i}\right|_{\lambda_i} = 0$$

(5.38)

$$2 \sum_j \lambda_j C(\mathbf{r}_i, \mathbf{r}_j) - 2c(\mathbf{r}_i, \mathbf{r}) = 0$$

(5.39)

From the last we can directly see

$$C(\mathbf{r}_i, \mathbf{r}_j)\lambda_j = c(\mathbf{r}_i, \mathbf{r})$$

(5.40)

A field that is completely described by second order statistics is Gaussian, since Kriging only uses the covariance function, hence the interpolated field must be Gaussian as well. Next we will give a formal derivation of this statement.
5.D.1 Gaussian Probability, alternative derivation

Another way of seeing that is by adopting a Bayesian viewpoint. We follow the line and notation of Rybicki & Press (1992). Let \( \tilde{f} \) be the estimated value then the augmented vector \( \tilde{f} \) is:

\[
\tilde{f} = \begin{pmatrix} f \\ \hat{f} \end{pmatrix}
\] (5.41)

Where \( \tilde{f} \) is a realisation of a Gaussian stochastic process, the elements \( f \) are already known and \( \hat{f} \) is modelled as a random deviate from the same stochastic process, i.e.

\[
P(\tilde{f}) \propto \exp\left[-\frac{1}{2} \tilde{f}^T C^{-1} \tilde{f}\right]
\] (5.42)

\( C \) is the covariance matrix of the stochastic process. Note that the connection to constrained random fields (Bertschinger 1987; Hoffman & Ribak 1991; Sheth 1995; Van de Weygaert & Bertschinger 1996) is obvious when one considers the measured values as the constraints. The Kriging interpolation value is nothing more than the mean field value at the position. Assuming that the augmented vector is a Gaussian realisation,

\[
P(\tilde{f}) \propto \exp\left[-\frac{1}{2} \tilde{f}^T \tilde{C}^{-1} \tilde{f}\right].
\] (5.43)

Here \( \tilde{C} \) is a new covariance matrix given by;

\[
\tilde{C} = \begin{pmatrix} C & c(r, \tilde{r}) \\ c(r, \tilde{r})^T & c(\tilde{r}, \tilde{r}) \end{pmatrix}
\] (5.44)

Here \( c(r_i, r_j) \) is the covariance function, usually only dependent on the distance of between \( r_i \) and \( r_j \), \( c(\tilde{r}, \tilde{r}) \) is the self variance term, in case of secondary stationary fields and without noise this terms vanishes. The conditional probability of for \( \tilde{f} \) given the known data values \( f \) is

\[
P(\tilde{f}|f) = \frac{P(\tilde{f}, f)}{P(f)} = \frac{\hat{P}(\tilde{f})}{P(f)}.
\] (5.45)

Writing out the terms of \( \tilde{C}^{-1} \) using block inversion,

\[
\begin{pmatrix} C^{-1} + C^{-1} c(\hat{c} - cC^{-1} c)^{-1} cC^{-1} & -C^{-1} c(\hat{c} - cC^{-1} c)^{-1} \\ -(\hat{c} - cC^{-1} c)^{-1} cC^{-1} & (\hat{c} - cC^{-1} c)^{-1} \end{pmatrix}
\] (5.46)

Let us define a scalar \( s \) and a vector \( y \).

\[
s = (\hat{c} - cC^{-1} c)^{-1}
\] (5.47)

\[
y = -s C^{-1} c
\] (5.48)

Then using this inverted matrix in eq. 5.43 then we can write out the terms of eq. 5.45 using the matrix

\[
\tilde{C}^{-1} = \begin{pmatrix} C^{-1} + \frac{1}{s} yy^T & y \\ y^T & s \end{pmatrix}
\] (5.49)

Carrying out the inner product,

\[
P(\tilde{f}|f) \propto \exp\left[-\frac{1}{2} \left( \tilde{f}^T C^{-1} \tilde{f} + \frac{1}{s} y^T y f \hat{f} + 2\tilde{f} y^T \hat{f} + s \hat{f}^2 \right) + \frac{1}{2} \tilde{f}^T \tilde{C}^{-1} \tilde{f}\right]
\] (5.50)
Which can be rewritten as

\[
P(\hat{f}|f) \propto \exp \left[ -s(\hat{f} + s^{-1}y)^2 \right] = \exp \left[ -\frac{(\hat{f} - c^T C^{-1} f)^2}{(c - cC^{-1}c)} \right]
\]  

(5.51)

So the best estimate for \( \hat{f} \) is when

\[
\hat{f} = c^T C^{-1} f.
\]  

(5.52)

This also happens to be the mean of the estimator distribution. Hence the Kriging estimate is a best linear unbiased estimator with an estimator variance \( \sigma_f^2 \) equal to

\[
\sigma_f^2 = (c - cC^{-1}c)
\]  

(5.53)

### 5.E Local Lognormal Kriging Algorithm

The Kriging interpolation method we use in this article we used the following method.

**Input:** Point sample \( r_i \) and interpolated positions \( \hat{r}_j \) (here a grid)

**Construct the Delaunay Triangulation of** \( r_i \)

**For each** \( r_i \) **calculate the density** \( f(r_i) \) **and** \( \hat{f}_i = \ln f(r_i) \).

**For each interpolating position** \( \hat{r}_j \) **do:**

1. Identify the Tetrahedron, \( T_j \) **in which** \( r_j \) **is located**
2. Obtain the set of vertices \( v_k \) **that are the neighbours of the four vertices of** \( T_j \) **(first order neighbourhood)**
3. **(If second order neighbourhood, than add the set of neighbour-vertices of** \( v_k \) **to** \( v_k \))
4. Calculate the Covariance matrix \( C(v_k, v_l) \) **and Covariance vector** \( c(v_k, \hat{r}_j) \)
5. Solve the linear system

$$C(v_k, v_l) \lambda_l = c(v_k, \hat{r}_j)$$  \hspace{1cm} (5.54)

6. Interpolate the density at \( \hat{r}_j \) with

$$\hat{f}(\hat{r}_j) = \sum_{k=0}^{N} \lambda_k f(v_k)$$  \hspace{1cm} (5.55)

Take the exponent of the interpolated values \( \hat{f}(\hat{r}_j) = \exp \hat{f}(\hat{r}_j) \)

end

5.F Sampling schema and artifacts

After we compute density estimates for each particle in the Delaunay tessellation we sample the density field inside a regular cubic grid. The most common sampling schema is one sampling point at the centre of each voxel. However, since the density field described by the tessellation is in principle not band-limited\(^8\) the net effect of this simple sampling schema is aliasing of the sampled density field (Shannon 1949; Unser 2000) (see Figure 5.31. In the DTFE method aliasing introduces large errors when the sampling point hits the interior of a tetrahedron located inside a local high density neighbourhood. The artificially highly dense or hot voxels introduced by aliasing can be difficult or even impossible to properly remove. Ideally in order to avoid aliasing one should apply a low-pass filter to the signal being sampled before the sampling procedure. In our particular case the spatial filtering must be applied directly on the Delaunay tessellation. This involves integrating the density field inside each voxel taking

---

\(^8\)In N-Body simulations it is in fact limited by the softening length.
into account the contribution of each tetrahedra that is contained in the voxel and its intersections with the voxel’s boundaries. The implementation of this approach is extremely laborious and will not be treated here. Instead we opted to remove the hot voxels after the sampling procedure.

There are several techniques to reduce aliasing once the image (or density field in our case) has been obtained (see Crow (1981) for a review). One of the most widely used techniques is to sample the signal at several points per voxel and perform some kind of average. We perform this by sampling either at several fixed or random positions inside the voxel, effectively oversampling inside a higher resolution grid (see figure 5.32). This has the advantage that in the case of voxels located inside large tetrahedra both the mean and median of the sampling points equals the mean of the density field inside the voxel (since the interpolation inside the tetrahedra is linear).

We will instead use either (a) the oversampling method on a regular grid with three times as much resolution per dimension as the original grid giving a total of 27 sampling points per voxel and assign their median value to the voxel. In the case of fields with large variations at small scales, such as the density field delineated by the Delaunay tessellations, the median is a more robust estimate than the mean. Alternatively (b) one may sample the voxel at random location within the voxel and take the mean of those values. The density fields here have been generated by oversampling each voxel approximately 50 times and taking the mean of those values.