DISSERTATIONES
ASTRONOMIAE
UNIVERSITATIS
TARTUENSIS
**22**

MAARJA BUSSOV

# Clustering Analysis for Astrophysical Structures

DISSERTATIONES ASTRONOMIAE UNIVERSITATIS TARTUENSIS
**22**

DISSERTATIONES ASTRONOMIAE UNIVERSITATIS TARTUENSIS
**22**

**MAARJA BUSSOV**

# Clustering Analysis for Astrophysical Structures

UNIVERSITY OF TARTU
Press

This study was carried out at the Tartu Observatory, University of Tartu, Estonia.

The Dissertation was admitted on November 9, 2020, in partial fulfilment of the requirements for the degree of Doctor of Philosophy in physics, and allowed for defence by the Council of the Institute of Physics, University of Tartu.

Supervisors: Prof. Elmo Tempel,
Tartu Observatory, University of Tartu
Tõravere, Estonia

Prof. Radu S. Stoica,
Université de Lorraine, CNRS, IECL
Nancy, France

Opponent: Dr. Pekka Heinämäki
Department of Physics and Astronomy, University of Turku
Turku, Finland

Defense: February 5, 2021, University of Tartu, Estonia

ISSN 1406-0302
ISBN 978-9949-03-531-1 (print)
ISBN 978-9949-03-532-8 (pdf)



University of Tartu Press
www.tyk.ee

# CONTENTS

## List of original publications

**This thesis is based on the following publications:**

I Tempel, E., Kipper, R., Saar, E., **Bussov, M.**, Hektor, A., Pelt, J. 2014, *Galaxy filaments as pearl necklaces*, Astronomy & Astrophysics, 572, A8.

II Tempel, E. & **Bussov, M.** 2014, *Filamentary pattern in the cosmic web: Galaxy filaments as pearl necklaces*, Proceedings of the International Astronomical Union, 11(S308), 236-241.

III **Kruuse, M.**, Tempel, E., Kipper, R., Stoica, R. S. 2019, *Photometric redshift galaxies as tracers of the filamentary network*, Astronomy & Astrophysics, 625, A130

IV **Bussov, M.** & Nättilä, J. 2020, *Segmentation of turbulent computational fluid dynamics simulations with unsupervised ensemble learning*, in preparation

**Other related publications of the dissertant:**

V **Kruuse, M.**, Tempel, E., Kipper, R., Stoica, R. S. 2019, *The bivariate J-function to analyse positive association between galaxies and galaxy filaments*, 51es Journées de Statistique 2019, Nancy, France

VI Tempel, E., **Kruuse, M.**, Kipper, R., Tuvikene, T., Sorce, J. G., Stoica, R. S. 2018, *Bayesian group finder based on marked point processes*, Astronomy & Astrophysics, 618, A81

**Author's contribution to the publications**

Author has made considerable contributions to the following original publications. The following list gives details on the author's work in each of the papers. The Roman numerals correspond to those in the list of publications.

**Publication I.** The results of this paper were first published in the master's thesis of the author, which was fully written and analysed by the author. She took part of the analysis in the publication and helped write the text of the published paper.

**Publication II.** The conference proceeding summed together the work done by the author in her master's thesis and the paper published by the author and co-authors. The author took part in the creation of this combined paper.

**Publication III.** The author performed all the necessary calculations of the analysis. She wrote the programming codes for the analysis and implemented them on the datasets of the study. The author outlined the paper structure, wrote the majority of the text and drew all the figures.

**Publication IV.** The author performed the analysis in collaboration with the co-author. She provided the idea for methodology choices to tackle the problem and carried out the mathematical development of the methodology. She wrote big parts of the text and carried out the data analysis in collaboration with the co-author.

**Publication V.** The author performed an additional analysis for the application of the clustering analysis method first published in Paper III. The proceeding summarizes the applicability of the methodology for astronomical data. All the necessary calculations of the analysis, drawn figures and the text is written by the author.

**Publication VI.** The author took part in the writing of the paper. The contribution is in the mathematical representation of the algorithm in the paper.

# INTRODUCTION

Data science has established itself as the fourth pillar of scientific discovery in addition to experimental, theoretical and computational science. This thesis addresses the use and further development of state-of-the-art data science methods for questions encountered in cosmology and astrophysics. The corresponding datasets are galaxy catalogues and turbulent magnetically dominated plasma simulation data. The data analysis problems for these seemingly different two datasets are after all similar: we wish to detect clustering of data points.

Galaxies are one of the largest objects in the Universe that are visible with amateur observational equipment. For over 50 years, systematic surveys viewing the cosmos in different wavelengths have pioneered in the discovery of new astrophysical objects and mapping the large scale of the Universe. Moreover, developments in observational technology and computer capabilities have created an abundance of astronomical data and thereby possibilities for applying complex analysis tools. In addition to advancements in observational and analysis methods, high-performance computing systems aka supercomputers have made it possible for astrophysicists to simulate physical phenomena, events otherwise not possible to be observed in such detail or abundance; or anymore in the cosmos. This thesis applies and develops clustering tools for observational archives and simulated data. The interest lies in detecting clustering in 1-dimensional to n-dimensional spaces.

Surveys have been the driving force of modern astronomy. The surge of astronomical data started in the late nineties with the first big ambitious astronomical surveys, such as the Two Micron All Sky Survey (2MASS) (Skrutskie et al. 2006), 2dF Galaxy Resdhift Survey (2dfGRS) (Colless et al. 2001) and Sloan Digital Sky Survey (SDSS) (York et al. 2000). In mapping the visible Universe, one of the pioneering surveys is the SDSS. Since the start of data collection in 2000, SDSS has gathered information about hundreds of millions of astronomical objects, some of them reaching as deep as redshift 6, when the Universe was less than $10^9$ years old. This archive of objects has offered more cosmological discoveries than any of its predecessor and continuing to be in the front line of science. The SDSS offers the most comprehensive database of observed galaxies. Maps created out of these galaxy catalogues manifest the web-like pattern of matter in the present-day Universe; revealing the large scale structure of the Universe, which consists of high-density galaxy clusters, bridging elongated galaxy filaments, 2-dimensional sheets of galaxies and almost empty voids. The previously described complex network of galaxies is clearly visible on Figure 1. It is a map of galaxies residing in the Universe, obtained from the Sloan Digital Sky Survey data release 8.

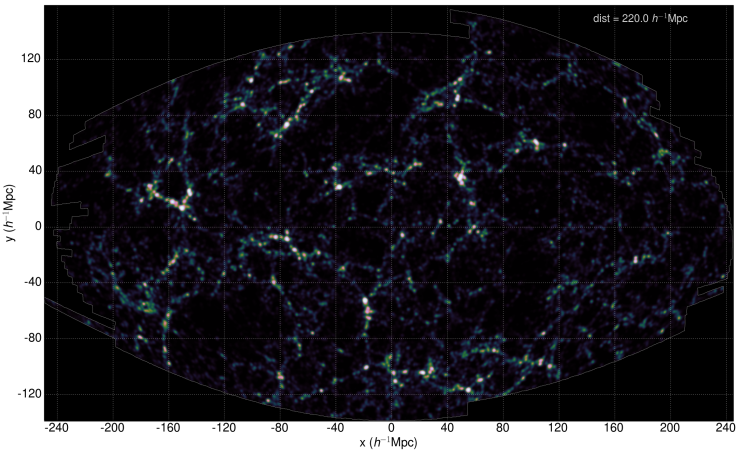



**Figure 1**: Large scale structure of the Universe, plotted is the luminosity density of galaxies. Image: L. J. Liivamägi.

The energy content of the Universe is 4.6% of baryonic matter, which is the matter that radiates in the electromagnetic spectrum of wavelengths, 24% of dark matter, which is only traceable by its gravitational effects, and 71.4% of dark energy. The structure of the Universe has been studied for decades (Sunyaev & Zeldovich 1970; Jõeveer et al. 1978; Peebles 1980; Bond et al. 1996). The cosmic web is the leading evidence for the existence of the dark matter, providing us with a view of its distribution in the Universe. Galaxy clusters consisting of hundreds of galaxies are held together by the gravitational pull of the dark matter. Galaxies in the filamentary structures are driven towards merger with the clusters by the same gravitational pull of dark matter collapsing towards the cluster. Galaxies inside the lower density sheets and voids are dragged into filamentary highways. Put shortly, matter in the Universe moves towards higher density regions, whilst dark energy is enlarging the distances between objects and tears gravitationally unbound objects forever further. Thus, analysing the galaxy distribution stands as one of the leading tools in understanding the dark matter. In this thesis we study the dark matter driven elongated galaxy filaments with techniques from spatial point pattern theory.

Computer-driven explorations in physics have emerged as an invaluable source for new scientific discoveries. Supercomputing has paved a way towards a surge of data in high-energy astrophysics, including astrophysical plasma simulations. In this thesis, we apply clustering methods to a plasma pool of magnetically dominated turbulent charged particles. Most interest lies in mapping particles belonging to thin elongated current sheets. These thin elongated structure elements are clearly visible as the darker blue or deeper red lines in the 3−dimensional Figure 2. After their de-

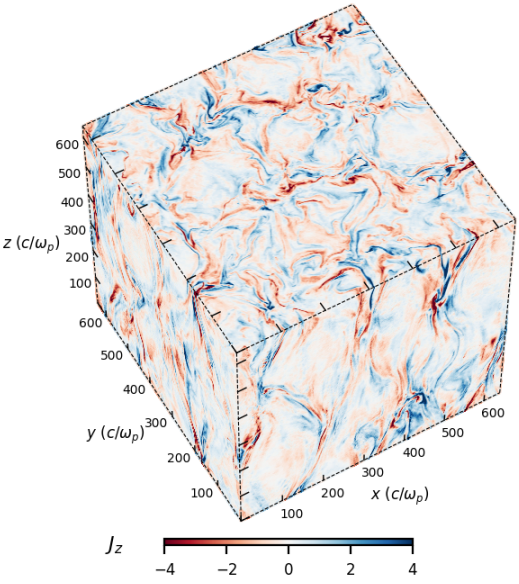



**Figure 2**: Computer simulation image of turbulent astrophysical plasma, the drawn physical quantity is the current $J$ in z-direction. Image: J. Nättilä.

tection from a pool of many millions of plasma particles, it is possible to study these intermittent structures in more detail. Understanding their dynamics is of great importance to the plasma astrophysics. Physically the current sheets are the main agent in turbulent plasma that accelerate non-thermal particles. In addition to astrophysical plasma, understanding the plasma sheets is key for plasma confinement needed in fusion reactors as well as for understanding space weather.

Different clustering analysis techniques are applied in this thesis in order to study observational galaxy datasets and astrophysical plasma simulation data, both consisting of millions of objects. We explore the galaxies residing in filaments to find patterns in their distances using spatial point patterns theory. Furthermore, we analyse possible clustering between two different types of catalogues of galaxies and a stochastically retrieved network of filamentary spines using advanced spatial point pattern techniques. This thesis applies spatial correlation and clustering algorithms to find correlations in a 1-dimensional pattern of galaxies, and detects a clustering signal on a region of a sphere between a spatial pattern of galaxies and a stochastically retrieved catalogue of objects.

For the astrophysical plasma simulations we will apply an unsupervised machine learning algorithm called the Self-Organizing Map to retrieve clusters from simu-

lation data and develop a stacking framework to improve the desired output. The proposed novel stacking algorithm for physical structure detection from astrophysical plasma simulation images increases the robustness of the detected clusters.

The thesis is structured as follows: The Chapter 1 gives an overview of spatial clustering algorithms, starting with a brief overview of the rigorous mathematical theory of spatial point processes, followed by the basic definitions of the field. Then, the spatial point pattern analysis tools are mathematically defined and their estimation functions retrieved. The Chapter 2 introduces the unsupervised clustering algorithms providing also an overview of the state-of-the-art in the rapidly developing field of computer vision algorithms used for image dissection. After that, the unsupervised learning algorithm Self-Organizing Map is mathematically defined. Last, the developments for obtaining more robust object boundaries for unsupervised image segmentation results are described. The Chapter 3 introduces the physical datasets analysed with the aforementioned algorithms, together with physical results from the analysis. First, the galaxy dataset are described and the obtained results from the analysis shown. Second, the astrophysical plasma simulation data is explained and the results from unsupervised clustering algorithms are highlighted. The thesis ends with a concluding Chapter 4 that provides a synthesis of the papers of this thesis.

# 1 Spatial clustering algorithms

This chapter introduces spatial point patterns as well as statistical tools that are used to study the clustering of spatial objects of interest in this thesis. First, a brief overview of spatial point pattern analysis is provided. Second, the mathematical definition of a point process with most important examples discussed. Then, the pair correlation function, which analyses spatial correlation of distances between pairs of points of a point pattern, is defined. The chapter discusses also the bivariate $J-$function, which studies nearest-neighbour distances from one spatial point pattern to another random set of spatial objects.

## 1.1 Overview of spatial point pattern analysis

The following brief overview of the spatial point pattern analysis is based on the presentation in the books Baddeley et al. (2015) and Illian et al. (2008). A spatial point pattern is an outcome of an experiment. The random mechanism that generates a spatial point pattern is referred to as a point process. If we generate the same random point processes multiple times under identical conditions, the outcomes differ each time.

Many observations can be represented as a spatial pattern of points. Already in the 19th century the epidemiologist John Snow mapped cholera victims residences and showed their proximity to a contaminated water pump (Snow 1855). More contemporary examples from other domains of sciences include the studies of urban areas (Brelsford et al. 2018; Huynh 2019); roads in a tropical forest (Kleinschroth et al. 2016); locations of trees in native woodland vegetation (Chang et al. 2013); multiple biological species in an observation area (de Jongh & van Lieshout 2020) or even instances of burglaries in the city of London (Povala et al. 2019). In astronomy, the spatial locations of galaxies in the Universe or stars in a globular cluster have been described by a configuration of points in an observation window for decades (Jõeveer et al. 1978; Martinez & Saar 2001; Kuhn et al. 2012).

These events and objects can be represented by their spatial locations in the observed space. Furthermore, each point in a spatial pattern of points can be assigned with additional characteristics, for example the luminosity and/or the type of the galaxy. These characteristics are called marks, and such a point pattern is referred to as the marked point pattern.

Exploratory analysis of spatial point patterns can reveal spatial correlations, trends in their spatial density, or even positive dependence to other spatial patterns. The latter can be observed if the spatial data is classified into different types, or we

are observing two patterns of points drawn from two processes. Analysing the spatial data with summary statistics gives estimates for the aforementioned features in the point pattern. In such cases, the different possible types of distances for the point pattern are calculated, e.g., the pair-wise distances among the points in the pattern, the smallest distance from a point of the pattern to another point of the pattern (nearest-neighbour distance), or the distance from a fixed point in the observation window to the first point of the point pattern (empty-space distance).

There exists also further possibilities for analysing a spatial point pattern by fitting a statistical model to the observed pattern. Constructing and fitting a point process model on point pattern data means synthesizing knowledge gained from a rigorous exploratory analysis. A statistical model is a more exhaustive portrayal of the point pattern. A good model is easy to interpret and simple; it contains information about the nature and spatial extent of the interactions and correlations present in the point pattern. In the case of a marked point pattern, the influence of the marks is explicit. Such models can be used to simulate the observed pattern, and with the possibility of changing parameters, new insight into the data can be obtained. Examples of statistical models include the Gibbs model, which was first designed to describe the physics of gas molecules, and marked space-time point process models, which can be used to analyse earthquakes (Ogata 1998). In the latter case, the process has in addition to its position $x$, marks of time $t$ and magnitude $m$.

Observational physics has developed and used broadly the tools of spatial statistics. Physical phenomena follow the laws of physics, which can be modelled with restrictions by statistical models. The structure features of the Universe referred to as galaxy filaments can be modelled from a spatial pattern of galaxies by an object point process with interactions (Stoica et al. 2010; Stoica 2010). Such a model is developed for 3−dimensional galaxy catalogues and referred to as the Bisous model in Tempel et al. (2014).

In this thesis, we will probe the structure of the Universe with tools from spatial statistics, with the main goal to find signals of spatial correlation and clustering, which indicates that the observed points tend to be closer together than would be expected for a completely random pattern. We will search for patterns in the galaxy locations in the modelled galaxy filaments, in Paper I and Paper II. Spatial clustering between the Bisous filaments and a new dataset of galaxies in the observed space of the Universe is discussed in Paper III.

## 1.2 Point processes

A point process is a random mechanism in an observed sampling window $W$ that produces a point pattern. In spatial statistics, we are interested in identifying trends in the

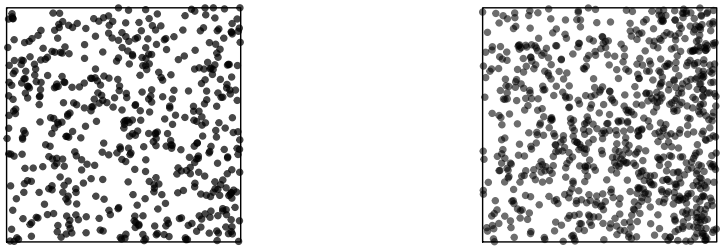


**Figure 1.1**: Two distinct simulations of the Poisson point process. The left panel shows the homogeneous Poisson process with intensity $\lambda = 500$ in a unit square. The right panel shows the inhomogeneous Poisson process with an intensity function dependent on the $x$ axis location.

spatial arrangement of these observed points. In this section, we define a point process and highlight most known point processes, following the presentations of Stoica (2010) and Baddeley et al. (2007). More comprehensive mathematical presentations of point processes can be found in the monographs of Stoyan & Stoyan (1994), van Lieshout (2000), Møller & Waagepetersen (2004), Illian et al. (2008), Chiu et al. (2013), and Baddeley et al. (2015).

Let $W \subset \mathbb{R}^2$ be a compact set and $\nu$ be the Lebesgue measure in $\mathbb{R}^2$ and $(W, \mathcal{B}_W, \nu)$ the natural restriction to $W$ of $(\mathbb{R}^2, \mathcal{B}, \nu)$, with $\mathcal{B}$ being the associate Borel $\sigma-$algebra. $W_n$ is the set of all unordered configurations $\mathbf{w} = \{w_1, w_2, ..., w_n\}$, which consist of $n$ not necessarily distinct points $w_i \in W$, for $n \in \mathbb{N}$. We consider the configuration space given by $\Omega_W = \bigcup_0^\infty W_n$ equipped with the $\sigma-$ algebra $\mathcal{F}_W$ generated by the mappings $\{w_1, ..., w_n\} \to \sum_{i=1}^n \mathbb{1}\{w_i \in B\}$ that count the number of Borel sets $B \in \mathcal{B}_W$. This leads us to define a point process: a point process on $W$ is a measurable map from a probability space $(\Omega_W, \mathcal{F}_W)$. In other words, a point process is a random configuration of the points lying in $W$. Simplest point processes are the uniformly random points and the binomial point process. A point process $X$ in the observation window $W$ with $n \in \mathbb{N}$ identically distributed points from a density function $f$ is referred to as the binomial point process, denoted as $X \sim \text{binomial}(W, n, f)$. For simulating the binomial point process, each point is generated independently by sampling its coordinates $(x, y)$ from a density function $f$ in the sampling window $W$.

The Poisson point process represents complete spatial randomness and is often used as a reference process in many spatial statistics tools. Let $\beta : W \to ]0, +\infty[$ be the intensity function of a Poisson point process in $W$. Then, a Poisson process is defined by the following two properties:

1. for any bounded set $B$, $N(B)$, the number of points in $B$ is a Poisson random variable with mean $\Lambda(B) = \int_B \beta(x)\mathrm{d}\nu(x)$,

2. for any $B_1, ..., B_m$ disjoint bounded sets, the point counts $N(B_1), ..., N(B_m)$ are independent random variables.

If the Poisson point process has a constant intensity $\beta$ = const, then the process is said to be homogeneous. If the average density of points $\beta$ is a function of spatial location $\beta(x, y)$, then the Poisson point process is inhomogeneous.

Figure 1.1 highlights the difference of the homogeneous and inhomogeneous Poisson processes. At any location on the left panel the density of points is visibly not dependent on the location inside the unit square. On the right panel on the other hand, the point density is clearly dependent on the $(x, y)$ location in the unit square. These theoretical processes carry as vital reference cases for many spatial statistics tools, such as the ones we will describe in the following sections.

## 1.3 Correlation and clustering analysis of point processes

### 1.3.1 The pair correlation function

A wide variety of techniques exist for studying dependencies in a point pattern. This section starts by describing the pair correlation function following the representations of Møller & Waagepetersen (2017), Baddeley et al. (2015), Martinez & Saar (2001) and Pons-Bordería et al. (1999). A more rigorous mathematical formalism for the pair-correlation function with examples to astronomical study cases is done in the paper of White (1979). The pair correlation function is easily interpretable and allows to study the correlation between points in a point pattern, here correlation is a summary statistic describing the association between points in a point pattern. The pair correlation function is widely known in the field of astronomy and was used for investigating galaxy-pair distances inside galaxy filaments in Paper I and Paper II.

The pair correlation function estimates the excess probability for two points lying at a distance $r$ from each other in comparison to complete spatial randomness (CSR). Its is a second order statistics for a point process. The intensity $\lambda$ of a point pattern is a first-order characteristic of the process. Let $\lambda^n$ be the $n$th order joint intensity function. Then for pair-wise distinct points $u_1, \ldots, u_n$ in the observation window $W$ with infinitesimally small volumes $\mathrm{d}u_1, \ldots, \mathrm{d}u_n$, we can define $\lambda^{(n)}(u_1, \ldots, u_n)\mathrm{d}u_1, \ldots, \mathrm{d}u_n$ as the probability of observing a point in each of the $n$ infinitesimal volumes. For a pair of points $u, v \in W$ the pair correlation function (pcf) is

$$g(u, v) = \frac{\lambda^{(2)}(u, v)}{\lambda(u)\lambda(v),} \tag{1.1}$$

where $u \neq v$. The case $g(u, v) > 1$ means that there is attraction between the points $u$ and $v$. The values $g(u, v) < 1$ indicate inhibition between $u$ and $v$; $g(u, v) = 1$ corresponds to $u$ and $v$ locating independently of each other. Let us denote with $r = d(u, v)$ the distance between the points according to the corresponding distance metric $d$. The pair correlation function can be expressed to depend only on the distance between the points if the studied point process is homogeneous (invariant under translations) and isotropic (invariant under rotations) (Møller & Toftaker 2014; Møller & Waagepetersen 2007). In such case, we can write the following representation for the pair correlation function

$$g(u, v) = \frac{\lambda^{(2)}(d(u, v))}{\lambda^2} = g(r). \tag{1.2}$$

The pair correlation function $g(r)$ and correlation function $\xi(r)$ have the following relationship

$$g(r) = \xi(r) + 1. \tag{1.3}$$

Equivalently to the discussion above, in case of complete spatial randomness, the pair correlation function in Equation (1.3) takes the value $g(r) = 1$. If the distance $r$ between the pairs of points of a process are less frequent than would be expected for uniformly distributed points, the pair correlation function will take the value $g(r) < 1$. In case of $g(r) > 1$, the interpoint distance $r$ is more frequent than expected under the hypothesis of a stationary Poisson point process.

The pair correlation function highlights the preferred and disfavoured distances between points in a point process in comparison to a completely random pattern of points. It is an easily interpretable statistical quantity used vastly in the field of astronomy (Davis et al. 1988; van de Weygaert 1991; Buryak & Doroshkevich 1996; Somerville et al. 1997). Baddeley et al. (2015) emphasises the difference between pair correlation function and correlation in statistical sense: the pair correlation function describes correlation with possible values from 0 to infinity, when the statistical correlation between two random variables can obtain values from $[-1, 1]$.

### 1.3.2 Estimation of the pair correlation function

There is abundant literature about the ways to estimate the correlation function $\xi(r)$ in Equation (1.3) for point processes (Chiu et al. 2013; Doguwa 1990; Fiksel 1988; Baxter & Rozo 2013). The following brief mathematical representation follows that of Pons-Bordería et al. (1999) and Davis & Peebles (1983).

We are observing a manifestation of a point process in an observation window $W$, with $N$ points. The goal of the analysis is to study possible spatial patterns in the locations of the points in $W$. This means that we are interested in studying whether

there exists preferred distances between the pairs of points. As the reference case randomly distributed points are generated, which describes complete spatial randomness in the positioning of points in the observation window $W$. The random catalogue consist of $N_{rd}$ random points. Most often, uniformly distributed points or a Poisson point process are generated.

To define the necessary metrics, we draw two circles of radius $r$ and $r+\Delta r$ centred on a point in the point pattern in the observation window $W$, $\Delta r$ being a very small increment of distance. Then we can define $DD(r)$ as the count of pairs of points of the studied point pattern in the observation window $W$ that lie at a distance $[r, r + \Delta r]$ from each other. $DR(r)$ is defined as the count of pairs of points of the studied point pattern, and the generated random point pattern in the observation window $W$ that lie at a distance $[r, r + \Delta r]$ from each other.

A wide variety of estimators for the correlation function $\xi(r)$ exist in literature (Davis & Peebles 1983; Landy & Szalay 1993; Hamilton 1993). The main differences between them is in the correction for edge effects (Pons-Bordería et al. 1999; Kerscher et al. 2000). In the following we provide a description for the most common correlation function estimator applied in Paper I and Paper II. The Davis & Peebles (Davis & Peebles 1983) correlation function estimator is

$$\hat{\xi}(r)_{\mathrm{DP}} = \frac{N_{\mathrm{rd}}}{N} \cdot \frac{DD(r)}{DR(r)} - 1 \tag{1.4}$$

The estimation for the pair correlation function $\hat{g}(r)$ is obtained by inserting Equation (1.4) in Equation (1.3). The previously mentioned estimator requires the studied point pattern to be isotropic, alternative estimators are proposed in Stoyan & Stoyan (1994). The estimation function $\hat{\xi}(r)_{DP}$ needs a big enough set of random points, in order to increase its accuracy; the size of which can be determined by the use of numerical tests. The choice of $\Delta r$ is arbitrary and most often determined by the scales of interest in the analysis, such that the smallest pair-wise distances are existent in the dataset. Pons-Bordería et al. (1999) and Kerscher et al. (2000) present a comprehensive study on the biases, dependences on random samples and the errors of different correlation function estimators. Pons-Bordería et al. (1999) finds no universal preferred estimator for all cases of study, the choice should be done dependent on the data set and analytical task at hand. On the other hand, Kerscher et al. (2000) attributes the Landy & Szalay (Landy & Szalay 1993) estimator to have the best performance.

In addition to the Davis & Peebles spatial correlation function estimator $\hat{\xi}(r)_{\mathrm{DP}}$ the border-corrected Landy–Szalay (Landy & Szalay 1993) and the Hamilton (Hamilton 1993) estimator are applied to the study of pair-wise distances between galaxies in a spatial pattern of galaxies in Paper I and Paper II. All these spatial correlation

techniques were able to reveal the preferred distances between galaxies inside galaxy filaments, unknown until then.

### 1.3.3 The bivariate $J-$function

The previously mentioned spatial correlation techniques study all pair-wise distances between distinct pairs of points of a point pattern, but additional information can be retrieved from patterns of points if only the nearest-neighbour distances are analysed. Such an analysis can be done with the $J-$function. It is a summary statistics technique, which analyses the positive association (clustering) of points in a point pattern. The method can be generalized to determine clustering between a point pattern and any set of random objects, yielding the bivariate $J-$function (Møller & Waagepetersen 2004; van Lieshout 2000). The bivariate $J-$function was until now a mostly unknown clustering analysis technique for astronomers. The tool was introduced to the community with the Paper III, where spacing between a galaxy dataset and filamentary pattern were analysed.

The bivariate $J-$function carries information about the dependence between these spatial patterns. Mathematically rigorous descriptions for the summary statistics described in the following can be found in van Lieshout (2000), Møller & Waagepetersen (2004), Illian et al. (2008), Chiu et al. (2013), and Baddeley et al. (2015). This sections follows the representations given in Baddeley et al. (2015) and Paper III.

We are observing two processes, a point process $X$ and a random set $Y$. To explore the spacings between a point pattern and any random set of objects with the use of the bivariate $J-$function, the empty-space function and the bivariate nearest-neighbour function need to be defined. Also we need to assume that $(X, Y)$ is jointly stationary (Foxall & Baddeley 2002) i.e. the distributions of the bivariate process $(X, Y)$ and of the bivariate process $(X + e, Y + e)$ are identical for any transition vector $e \in \mathbb{R}^3$.

The empty-space function $F_Y(r)$ is a cumulative distribution function of the distances from an arbitrary location $u$ in the observation window $W$ to the nearest random object of $Y$:

$$F_Y(r) = \mathbb{P}\{d(u, Y) \leq r\}. \tag{1.5}$$

The bivariate nearest-neighbour function $G_{X,Y}(r)$ is the cumulative distribution function of the distance from a typical point of the point process $X$ to the nearest objects of the random set $Y$

$$G_{X,Y}(r) = \mathbb{P}\{d(u, Y \setminus u) \leq r \mid X \text{ has a point at u}\}. \tag{1.6}$$

Now we can define the bivariate $J-$function as

$$J_{X,Y}(r) = \frac{1 - G_{X,Y}(r)}{1 - F_Y(r)} \tag{1.7}$$

for all $r \geq 0$ such that $F_Y(r) < 1$. The $J_{X,Y}$ function measures association between two spatial patterns $X$ and $Y$ drawn from different processes (Foxall & Baddeley 2002). In case of a point pattern generated by the stationary Poisson point process with intensity $\beta$ describing complete spatial randomness (see Figure 1.1), the aforementioned summary statistics have exact formulas

$$\begin{aligned} F(r) &= 1 - \exp[-\beta\pi r^2], \\ G(r) &= F(r), \\ J(r) &= 1. \end{aligned} \tag{1.8}$$

If $X$ and $Y$ are independent, then $G_{X,Y} = F_Y$ and $J_{X,Y} = 1$. Values for $J_{X,Y}$ higher than 1 suggest negative association or "repulsion", and values lower than 1 suggest positive association or "clustering". The $J-$function can be interpreted as the ratio of two survival functions $1 - G_{X,Y}(r)$ and $1 - F_Y(r)$, which describe the distribution of distances of two different regimes (van Lieshout & Baddeley 1996; Foxall & Baddeley 2002). The numerator is the survival function of distances from a point of the process $X$ to the nearest object in the random set $Y$. The denominator is the survival function of distances from a fixed arbitrary point to the nearest object in the random set $Y$. The values of $J(r) < 1$ indicate that the survival function of the distances from a point of the process $X$ to the nearest object in the random set $Y$ is smaller than the survival function for the distance from a fixed arbitrary point to the nearest object in the random set $Y$. This indicates clustering between the patterns $X$ and $Y$. For the values of $J(r) > 1$ the survival function for the distance from a fixed arbitrary point to the nearest object in the random set $Y$ is smaller. This indicates repulsion between the patterns $X$ and $Y$.

### 1.3.4 Estimation of the bivariate $J-$ function

We are observing processes in a bounded window, which introduces edge effects. Thus we adopted the border-corrected estimation for the analysis as in van Lieshout & Baddeley (1996) and Foxall & Baddeley (2002). The shortest distances from a point $w \in W$ to a subset $A \subset W$,

$$d(w, A) = \inf_{a \in A} \|w - a\|. \tag{1.9}$$

The border-corrected estimator for the empty space function is

$$\widehat{F}_Y(r) = \frac{\sum_i \mathbb{1}\{d(w_i, W^c) \geq r\}\mathbb{1}\{d(w_i, Y) \leq r\}}{\sum_i \mathbb{1}\{d(w_i, W^c) \geq r\}}, \tag{1.10}$$

with $W^c$ the border of $W$ and $\{w_i, i = 1, 2, \ldots\}$ a finite family of arbitrary points in $W$. The border-corrected estimator of the bivariate nearest-neighbour function is

$$\widehat{G}_{X,Y}(r) = \frac{\sum_i \mathbb{1}\{d(x_i, W^c) \geq r\}\mathbb{1}\{d(x_i, Y) \leq r\}}{\sum_i \mathbb{1}\{d(x_i, W^c) \geq r\}}, \tag{1.11}$$

where $\{x_i, i = 1, \ldots\}$ is the observed finite-point configuration of $X$ and hence the border-corrected estimator for bivariate $J$ function is

$$\widehat{J}_{X,Y}(r) = \frac{1 - \widehat{G}_{X,Y}(r)}{1 - \widehat{F}_Y(r)}. \tag{1.12}$$

An approximation of the variance and other properties of the bivariate $J-$function estimator in Equation (1.12) is given in van Lieshout & Baddeley (1996, 1999); Foxall & Baddeley (2002).

# 2 Unsupervised clustering algorithms

This chapter introduces machine learning algorithms that are used for object dissection from images and 3D videos. Because of the scope of this thesis, we will focus on image segmentation using unsupervised clustering algorithm tools. First, we present a general overview of available tools and the state-of-the-art in image segmentation. Then the chapter describes more thoroughly an artificial neural network algorithm Self-Organizing Map (SOM). Last, a novel stacking algorithm developed in Paper IV for obtaining more robust boundaries for objects on an image is presented.

## 2.1 Overview of machine learning algorithms for image segmentation

Machine learning algorithms are utilized in a wide variety of concepts. In statistical modelling, data is described with models, which take into account assumptions about the data. Often the goal is to test a hypothesis about the data. The goal of machine learning is to make predictions about novel data with an algorithm that is derived from previous observations. In addition, these algorithms do not need explicit assumptions for the data.

Computer vision is maybe one of the most popular machine learning domains. The best algorithms dissecting objects from images or 3D videos include the non-local neural network algorithms proposed in Wang et al. (2017) and Zhu et al. (2019). Valada et al. (2019) forms a Self-Supervised Model Adaptation for recalibrating and fusing modality-specific feature maps, as feature vectors are often not invariant. Convolutional neural networks can be applied to goals other than image processing. Landrieu & Simonovsky (2018) improves semantic segmentation for large 3D point clouds. The point cloud is represented as interconnected shapes, which are then feed into a deep learning algorithm based on graph convolutions. These best-performing image and 3D point cloud processing tools are fully supervised learning algorithms that come with high computational cost. In addition, the algorithms need to be trained with vast amounts of pixel-level labelled data, which is not always available and wears down the analysis.

Convolutional neural networks, whose different architectures are the backbones of deep learning algorithms, are also shown to produce significant results with labels only on the image level or with bounding box information (Zhou et al. 2016). Such deep learning tools are referred to as semi-supervised learning algorithms.

If there is no classification information about the images available, the analysis can be done with unsupervised learning tools. For these algorithms only information of the input variables is known, but the aspired output is unknown. The algorithm will

learn a representation of the input data with a few explicit assumptions. One of the most known unsupervised clustering algorithms is the k-means (Lloyd 1982; Steinhaus 1956). More complex clustering analysis paradigms like Neural Gas (Martinetz & Schulten 1994) and Self-Organizing Map (Kohonen & Mäkisara 1989; Kohonen 2001) create an elastic cloud or a layer of neurons, which learn a 2−dimensional representation of the input data. Often their accuracy is not as good as fully supervised counterparts. Choosing a machine learning algorithm is highly dependent on the dataset being studied, possibility of labelling data and available computational resources.

To obtain insight about clusters representing physical phenomena in astrophysical simulation images, we wanted to train a machine learning model with the physical data. Options for that were linear regression analysis; prototype and distance based methods like k-means or learning vector quantization methods; linear threshold classifiers such as support vector machines; or non-linear regression such as deep neural networks. The choice of method is delicate and there exists an abundance of literature about basic models and their varied developments e.g., Ding & Hua (2014); Gatys et al. (2015); Schneider et al. (2009); Mwebaze et al. (2015); Basheer & Hajmeer (2000). The basic approach is to start simple and try to obtain insight of the data and its variation. Interpretation is key for each analysis tool, simple models being easier to understand and interpret. Results obtained should be reliable and applicable to new data. Another important step in the model design is the selection of features, which should represent the variability in the data. Feature architecture can be done by designed models. For example, the principal component analysis is a model designed for creating linear components of original data features (Cross 2015). These linear components describe the most of variability in the dataset and can be used as the features in the model training. Multiple learning algorithms are also designed for data visualization, such as the principal component analysis and artificial neural networks models like Self-Organizing Map. Preliminary analysis with data visualization tools bring insight into the data and simplify the further analysis process.

Image segmentation is labelling every pixel on the image with a corresponding class. In Paper IV we tackled the issue of image segmentation without knowledge about the ground truth of the pixels. This is the case in many astrophysical simulations and observational datasets. We trained a model, which is able to classify N-dimensional data pixels into clusters, where each cluster corresponds to a physical structure in an astrophysical plasma simulation. We obtained segmented pixels of the astrophysical simulation images, where each segmented structure corresponded to a physical phenomenon in the turbulent plasma. The structures are fine and their geometrical dimensions are of great interest to the astrophysics community.

The segmentation from image data is done in three steps:

1. An artificial neural network model is applied to obtain a set of cluster IDs.
2. The algorithm is applied multiple times independently on the same physical data, acquiring multiple sets of independent realisations for the cluster IDs.
3. The independent sets of cluster IDs are then stacked in a statistical framework as discussed in Section 2.3.

The aforementioned process is computationally very powerful due to trivial parallelization over independent inference steps.

## 2.2 Self-Organizing Map

A Self-Organizing Map (SOM), also known as Kohonen's map is a machine learning algorithm that has gained popularity as a clustering analysis method and as a tool for visualization in exploratory data analysis. The SOM belongs to a class of unsupervised algorithms that do not need a-priori knowledge of the different clusters; the method performs the separation of data into multiple clusters on its own. A rigorous mathematical presentation of the Self-Organizing Map is given in the monograph Kohonen (2001). In the following we will give a description of the SOM algorithm using Kohonen (2013), Kohonen & Mäkisara (1989) and Kohonen (2001).

A neuron is a data point assigned with an initial vector, which "lives" in the same dimension as the studied input sample vector. The neuron vector can be initially given random values sampled from the input sample space or more sophisticated starting initialisation can be done, if needed. Data analysis based initialization methods are discussed in Kohonen (2013) and Valova et al. (2013). A neural network is a layer of such neurons. In our study, the neurons in a Self-Organizing Map are positioned on a $2-$dimensional elastic grid, in such manner that the distances between the neurons indicate their vector distances. Figure 2.1 shows an example of a square shaped Self-Organizing Map (Kohonen's map) of neurons and the input sample vector.

Mapping observations of a dataset into an artificial grid, or into a number of generated centroids in the sample space, is one of the main assignments of unsupervised machine learning techniques. The SOM algorithm moulds an elastic network of neurons with the use of learning rules, neighbourhood information and competitive learning to represent the input data. The neurons on the 2-dimensional network can be combined to form clusters with the use of a distance metric. Each cluster is represented by a centroid, which describes the most of the variability in the observed cluster, and each input data vector is assigned to a cluster by mapping it to the most similar neuron on the neural map. The result of the chosen SOM is a

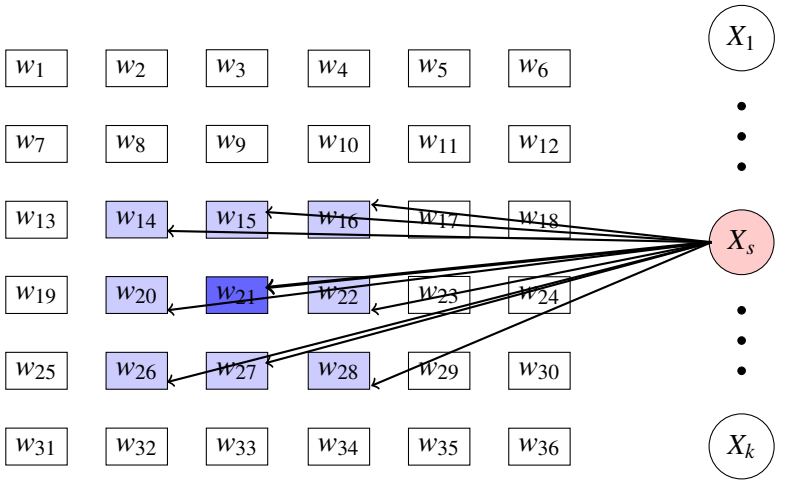



**Figure 2.1**: Kohonen's 2-dimensional rectangular map of neurons. Darker blue neuron on the map indicates the winning neuron, or the Best Matching Unit $w_{\text{BMU}}$ (see Equation (2.1)) for the input variable vector $X_s$. Light blue neurons indicate all the neurons in the neighborhood of the winning neuron ($N_{\text{BMU}}$). The arrows describe the competitive learning step (see Equation (2.2)).

2−dimensional map, where the dimensions describe most of the variability in the input data. A trained neural map encapsulates a 2−dimensional representation of the $l$−dimensional input space. At this step the neuron map does not represent clusters yet. The clusters are combined dependent on a smoothing length and the similarity measure between neurons.

Let us introduce the SOM method with some more mathematical rigour. Denote with $X = \{X_1, \ldots, X_k\}$ the input data vector, where each element $X_i$ describes a set of input variables $\xi_j$, such that $X_i = [\xi_1, \ldots, \xi_l] \in \mathcal{R}^l$, $\forall i \in \{1, \ldots, k\}$. We adopt a regular 2-dimensional rectangular-shaped Kohonen neural map of dimensions $(m, n)$. In passing, we note that a wide variety of different geometric shapes are possible, such as square, rectangle and hexagon. In addition to regular arrays, cyclic or growing networks can also be chosen. An accurate choice for the dimensions and architecture of the map will result in a faster convergence. Kohonen (2013) recommends selecting the map dimensions such that they describe the lengths of the first principal components and advises for a bigger map size to detect fine structures.

The distance metric used to compute the distance (or similarity) between an input variable vector and the neuron parametric vector is noted with $d(X_s, w_i)$. The Best Matching Unit, $w_{\text{BMU}}$ (BMU) from the neuron map, is determined by

$$w_{\text{BMU}} = \text{argmin}_i\{d(X_s, w_i)\} \tag{2.1}$$

$w_{\text{BMU}}$ represents a neuron on the 2-dimensional grid, which is the most similar to the sampled data vector $X_s$ according to the chosen distance metric (see Figure 2.1).

The competitive learning step to modify the neuron weight vectors in an SOM algorithm is given as

$$w_i(t+1) = \begin{cases} w_i(t) + \alpha(t)h_{ci}(t)[X_s(t) - w_i(t)], & w_i \in N_{\text{BMU}}, \\ w_i(t), & w_i \notin N_{\text{BMU}}, \end{cases} \tag{2.2}$$

where $t = 0, 1, 2, \ldots$ is a discrete time value. Here, $N_{\text{BMU}}$ is the neighborhood of the node $w_{\text{BMU}}$ on the neuron map, which consists of all the nodes up to a certain geometric distance $r$ from $w_{\text{BMU}}$. $N_{\text{BMU}}$ influences the map's ability to order itself and learn the underlying data distribution (Kohonen 2001; Lee & Verleysen 2002). In the SOM algorithm learning step, both the BMU and its spatial neighbors $N_{BMU}$ learn from the input vector (see Figure 2.1). These local interactions between neurons create a kind of elasticity for the map. The set of these neurons usually shrinks with time and is determined by the neighborhood function $h_{ci}(t)$. The $h_{ci}(t)$ acts like a smoothing kernel, featuring convergence $h_{ci}(t) \to 0$ as $t \to \infty$, determining the rate of change for neurons on the map. Details and different shapes for the neighborhood function are proposed in Kohonen (2001), Kohonen (2013) and Stefanovič & Kurasova (2011). $0 < \alpha(t) < 1$ is the learning rate, which determines the statistical accuracy of the neuron map and the ordering of the neurons on the map. It is suggested to normalize the sample vectors and the neuron vectors before the matching law (Equation (2.1)) and the learning step (Equation (2.2)) are applied. This will ensure that the neuron vectors and data sample have the same dynamical range.

Learning is a stochastic process, meaning that the accuracy of the mapping depends on the number of iteration steps. Kohonen (2001) recommends 500 times the number of network units on the neural layer. It is worth noting that the size of the input sample has no considerable effect on the number of needed iteration steps. The SOM algorithm is computationally inexpensive and thus high numbers of iteration steps can easily be tested. In the aforementioned representation we used the Euclidean distance metric, but other metrics may also be applied, as long as the matching rule in Equation (2.1) and the updating law in Equation (2.2) are mutually compatible. Kohonen (2001) highlights that the choices for all the aforementioned functions and their parameters are mainly chosen by trial and error.

The traditional SOM algorithm is highly flexible and can be modified in various ways. For example, Lee & Verleysen (2002) proposed an extension of a recursive neighborhood function to the classical algorithm that produces statistics to compare the performances of the algorithm. In Kohonen & Somervuo (2002) the algorithm was reinforced for clustering and visualization analysis of symbol sequences. In Somervuo & Kohonen (1999) the algorithm was associated with an entire feature

vector sequence and applied for speech recognition. Kohonen (1999) developed a fast evolutionary learning step that is based on the batch-type SOM. Hammer et al. (2004) presents an overview of several Self-organizing models, such as the temporal Kohonen map and a recursive SOM algorithm. The SOM algorithm has been used and further developed in thousands of scientific research papers and used in many fields of science including medical sciences, financial sciences, and speech analysis. An exhaustive list of the main research areas is given in Kohonen (2013).

## 2.3 The Statistically-Combined Ensemble framework

### 2.3.1 About the Statistically Combined Ensemble framework

The Statistically-Combined Ensemble (SCE) framework was derived in Paper IV. For the SCE a collection of evaluations from a clustering algorithm are added together to form an ensemble of cluster maps. The framework focuses on the unsupervised semantic segmentation of images and proposes metrics to stack independent realizations of the classified pixels. These are computationally fast combination operations for adding different image cluster matrices together. The SCE computes robust and accurate region of interest boundaries for the different pixel clusters by statistically averaging the segments over multiple clustering realizations.

### 2.3.2 Mask of a cluster

An unsupervised clustering algorithm applied to an image of size $r \times t$ will divide the original data into $n$ clusters. Each pixel $p_{ij}$, where $i \in \{1, \ldots, r\}$ and $j \in \{1, \ldots, t\}$, on this image is assigned to a cluster from the set of detected $n$ clusters $\{C_1, \ldots, C_n\}$, where $n, C_1, C_n \in \mathbb{N}$.

A mask $M_k$ is defined for every cluster $C_k$ in the set of clusters $\{C_1, \ldots, C_n\}$. A mask is a boolean matrix $M_k = (m_{i,j}) : r \times t$, such that

$$m_{ij} = \begin{cases} 0, & p_{ij} \notin C_k \\ 1, & p_{ij} \in C_k \end{cases} \tag{2.3}$$

Possible values in a mask matrix $M_k$ are $\{0, 1\}$. The element $m_{ij}$ in the mask matrix $M_k$ obtains value 1 if the pixel $p_{ij}$ is assigned to the cluster $C_k$ by the observed clustering algorithm and 0 otherwise. We define $n$ mask matrices $M_1, \ldots, M_n : r \times t$, one corresponding to each cluster in the set $\{C_1, \ldots, C_n\}$. The mask matrix is a simplified view of a cluster, describing the locations and area of the structure on the image. It provides a mapping from the cluster groupings into the initial data view.

The unsupervised clustering algorithm is applied $N$ independent times on the same image data. As a results each pixel $p_{ij}$ on the original image will have

$N$ independent cluster evaluations. This means that $N$ cluster sets are obtained: $\{C_1, \ldots, C_{n_1}\}, \ldots, \{C_1, \ldots, C_{n_N}\}$, with correspondingly $n_1, \ldots n_N$ elements residing in them. Using each of these cluster sets, we create $N$ independent sets of mask matrices $\mathcal{M} = \left\{\{M_1^1, \ldots, M_{n_1}^1\}, \ldots, \{M_1^N, \ldots, M_{n_N}^N\}\right\}$ according to Equation (2.3). We use a notation where the upper index of a mask matrix refers to the SOM instance index and the lower index to the detected cluster in that instance.

### 2.3.3 Mask to mask stacking

A set of masks $\mathcal{M}^b = \{M_1^b, \ldots, M_{n_b}^b\}$, $1 \leq b \leq N$, is randomly chosen from the set of all independent mask sets $\mathcal{M}$. This mask set $\mathcal{M}^b$ is named the base mask set. The framework will compare each mask in $\mathcal{M}^b$ to every other mask in the set $\mathcal{M} \setminus \mathcal{M}^b$. For every mask in $\mathcal{M}^b$ we acquire $n_1 + n_2 + \ldots + n_N - n_b$ comparisons. The comparison is done between every mask in a randomly chosen base mask set against every other mask set. This will be done as long as each set of masks has been chosen as the base mask set.

These comparison operations between masks will rely on a set theoretical base. We first calculate the intersection, union and sum matrices of the two masks that are being compared. We then combine these simple quantities to derive more complicated estimates to evaluate the goodness of fit between compared masks. In supervised segmentation the obtained result is compared with similarity measures to the known truth about the data. Whereas in the unsupervised learning case the ground truth does not exist. Therefore in this work we derived metrics to compare independent cluster evaluations for the same image data. These metrics are also combined and used for stacking of independent realizations on a chosen base mask. This created robust regions of interest for the detected physical structures on the images.

#### Similarity measures

For a mask $M_e^b$ in the base mask set $\mathcal{M}^b$, with $n_b$ masks and $1 \leq e \leq n_b$, and a mask $M_f^c$ from a mask set $\mathcal{M}^c$, with $n_c$ masks, $1 \leq f \leq n_c$ and $b \neq c$, we can calculate three different similarity measure matrices. These are: The union matrix $U : r \times t$, with possible values of $\{0, 1\}$, is defined as

$$U(i,j) = \begin{cases} 0, & M_e^b(i,j) = 0 \wedge M_f^c(i,j) = 0 \\ 1, & \text{otherwise} \end{cases} \tag{2.4}$$

where $i \in \{1, \ldots, r\}$ and $j \in \{1, \ldots, t\}$. The intersection matrix $I : r \times t$, with possible values of $\{0, 1\}$, is defined as

$$I(i,j) = \begin{cases} 1, & M_e^b(i,j) = 1 \wedge M_f^c(i,j) = 1 \\ 0, & \text{otherwise} \end{cases} \tag{2.5}$$

where $i \in \{1, \ldots, r\}$ and $j \in \{1, \ldots, t\}$. The sum matrix $R : r \times t$, with possible values of $\{0, 1, 2\}$, is defined as

$$R(i,j) = \begin{cases} 0, & M_e^b(i,j) = 0 \wedge M_f^c(i,j) = 0 \\ 1, & M_e^b(i,j) = 1 \wedge M_f^c(i,j) = 0 \quad \vee \\ & M_e^b(i,j) = 0 \wedge M_f^c(i,j) = 1 \\ 2, & M_e^b(i,j) = 1 \wedge M_f^c(i,j) = 1, \end{cases} \tag{2.6}$$

where $i \in \{1, \ldots, r\}$ and $j \in \{1, \ldots, t\}$.

These similarity measures are calculated between mask $M_e^b$ in the base mask set $\mathcal{M}^b$ and every mask in the set $\mathcal{M} \setminus \mathcal{M}^b$. This step is repeated until all masks in the set $\mathcal{M}$ have been chosen as a base mask set.

**Signal strength**

Using the matrices defined in Equations (2.4) and (2.5) a signal strength $s_I$ is defined. For any base mask matrix $M_e^b$ and a random mask $M_f^c$, where $b \neq c$, the signal strength scalar $s_I$ is

$$s_I = \frac{\sum_{i,j=1}^{r \times t} I(i,j)}{\sum_{i,j=1}^{r \times t} U(i,j)}. \tag{2.7}$$

The $s_I$ is a comparison measure, which estimates how well the chosen base mask cluster resembles other independent evaluations of clusters on the image. We denote two independent cluster evaluations identical if their mask matrices are identical. The signal strength gives an estimate for how similar is the intersection of two mask matrices to the union of the two masks. We note that $s_I$ is defined for the comparison of every mask in a base mask set to every other mask in an independent set.

If $\sum_{i,j=1}^{r \times t} I(i,j) \to \sum_{i,j=1}^{r \times t} U(i,j)$, then $s_I \to 1$. This means that this specific cluster has been detected by these two independent clustering runs to be at the same location and with similar projected shape on the image. Thus the two masks have the value 1 in the same locations and in same quantity in their $r \times t$ mask matrices. This indicates that the pixels are accurately classified by the observed masks.

The signal strength metric $s_I$ is identical to the Mean Intersection-over-Union (MIoU) metric, commonly used in supervised segmentation, with the difference that the MIoU metric is defined to compare the detected segmented objects to the ground truth — information that is lacking in the unsupervised learning case. In our case the $s_I$ metric is defined between independent instances of unsupervised clustering algorithms. It compares a cluster realization mask matrix to another independent evaluation of a cluster.

**Quality of cluster unions**

An accompanying measure to $s_I$ is the so-called quality measure $q_U$, which is constructed using the matrices defined in Equations (2.4), (2.5) and (2.6). The quality measure $q_U$ gives an estimate to the size of the symmetric difference of the two mask matrices. For any base mask matrix $M_e^b$ and a random mask $M_f^c$, where $b \neq c$, the scalar $q_U$ is defined as

$$q_U = \frac{\sum_{i,j=1}^{r\times t} U(i,j)}{\sum_{i,j=1}^{r\times t} R(i,j)} - \frac{\sum_{i,j=1}^{r\times t} I(i,j)}{\sum_{i,j=1}^{r\times t} R(i,j)}. \tag{2.8}$$

If $\sum_{i,j=1}^{r\times t} U(i,j) \rightarrow \sum_{i,j=1}^{r\times t} I(i,j)$ then $q_U \rightarrow 0$, meaning that the two independent stacked masks $M_e^b$ and $M_f^c$ have detected exactly the same structure in the same locations on the image. This highlights the best matching cluster masks, as they have a small residual area between their masks matrices (area of the symmetric difference in set theoretical notion). On the other hand as $\sum_{i,j=1}^{r\times t} U(i,j) \rightarrow \sum_{i,j=1}^{r\times t} R(i,j)$ then $q_U \rightarrow 1$, indicating that the intersection of the two masks is negligible and the cluster masks highlight different structures on the image. It describes the number of differently classified pixels compared to the sum of the pixels in the two cluster masks.

The Dice coefficient is a metric often used in supervised image segmentation. Using the intersection matrix (Equation (2.5)) and the sum matrix (Equation (2.6)) the Dice coefficient is

$$D = \frac{2 \cdot \sum_{i,j=1}^{r\times t} I(i,j)}{\sum_{i,j=1}^{r\times t} R(i,j)} \tag{2.9}$$

It is a metric positively correlated to the MIoU. The $q_U$ and $s_I$ are negatively correlated measures. The value $q_U \rightarrow 1$ as the intersection of the compared masks decreases. The value $s_I \rightarrow 1$ as the intersection of the masks increases. The quality of cluster unions $q_U$ metric is similar, but not equivalent, to the Dice coefficient.

### 2.3.4 Stacking multiple masks

Until now the metrics were created to compare two mask matrices from independent clustering evaluations. This concept will be now generalized to quantities able to average over the whole set of independent mask matrices in the set $\mathcal{M}$. A matrix $G$ is derived, which combines the signal strength $s_I$ and the quality of union $q_U$. It estimates the goodness of the fit between independent cluster masks. All comparisons for a base mask can be stacked to form a sum matrix $G_{\text{sum}}$. This measure quantifies the goodness of fit between all the independently detected cluster masks.

If multiple independent clustering algorithms have detected the same cluster indicating the same structure on the image, then these mask matrices will fit together

well. They will contribute to the sum matrix $G_{\text{sum}}$ for a base mask detecting the same structure. On the contrary, if a cluster mask represents a different physical structure on the image, it does not contribute significantly to the total integrated goodness measure.

A mask $M_e^b$ in the base mask set $\mathcal{M}^b$ is combined with every mask in the set $\mathcal{M} \setminus \mathcal{M}^b$. The goodness of fit of a cluster mask $M_e^b$ to any other cluster mask $M_f^c$ from $\mathcal{M} \setminus \mathcal{M}^b$ is defined as

$$G = \frac{s_I}{q_U} \cdot (M_e^b \cup M_f^c), \tag{2.10}$$

where $s_I$ and $q_U$ are the signal strength scalar and quality of union scalar of the masks $M_e^b$ and $M_f^c$. For the observed mask $M_e^b$ in the base mask set $\mathcal{M}^b$ we obtain $n_1 + n_2 + \ldots + n_N - n_b$ matrices according to the Equation (2.10), which have the corresponding quotient value in the union of the two compared masks. As the mask $M_e^b$ will have $n_1 + n_2 + \ldots + n_N - n_b$ signal strengths $s_I$ and qualities of union $q_U$. Then for every base mask $M_e^b$ in $\mathcal{M}^b$ all $n_1 + n_2 + \ldots + n_N - n_b$ of its $G-$matrices can be summed together to yield

$$G_{\text{sum}} = \sum_{k=1}^{N} \sum_{l=1}^{n_k} \Big( \frac{s_I}{q_U} \cdot (M_e^b \cup M_l^k) \Big), \tag{2.11}$$

where $k \neq b$. All mask sets in $\mathcal{M}$ will be chosen as a base mask set, acquiring $n_1 + n_2 + \ldots + n_N$ sum matrices $G_{\text{sum}}$, each of which characterizes the fit of a cluster mask to all other cluster masks detected in other independent cluster realizations.

The $G_{\text{sum}}$ for a base mask is constructed such that the value added to the union of the two masks will be high for pairs that have detected similar structures in the image. This means they have a similar number of pixels in same locations assigned to the detected cluster and thus their mask matrices have value 1 in same locations. Mask pairs that detect distinct structures on the image contribute a negligible amount to the $G_{\text{sum}}$, since the $s_I/q_U$ will be close to zero in value for those cases.

In Paper IV we developed and applied the stacking framework on clusters detected by the Self-Organizing Map algorithm from astrophysical plasma simulation images. We applied the unsupervised clustering algorithm to the image data $N$ times, which resulted in $n_1 + n_2 + \ldots n_N$ cluster masks. Then $N$ $G_{\text{sum}}$ matrices were derived. Each pixel in the original image view of a detected cluster obtains a value between 0 and 1 illustrating its stability of belonging to the cluster with that shape and location.

Similarly to the matrix representation of the quotients for each mask comparison a scalar value for the goodness of fit for a base mask $\mathcal{M}_e^b$ can be defined:

$$g_{\text{sum}} = \sum_{\mathcal{M} \setminus \mathcal{M}^b} (\frac{s_I}{q_U}), \tag{2.12}$$

where the sum is over all the comparisons between $\mathcal{M}_e^b$ and every mask other than $\mathcal{M}^b$ in $\mathcal{M}$. In Equation (2.12) all the quotients of the signal strength $s_I$ and quality of union $q_U$ are added together, yielding a scalar value for the base mask. By ordering the scalar $g_{\text{sum}}$ values of all base masks, we can detect the base mask that fits the best with all the other independent realizations of that cluster. That base mask has learned most of the information that describes the cluster. The independent base masks classifying the same cluster will likely follow the winning base mask in the sequence of the ordered $g_{\text{sum}}$ values. In the study of unsupervised ensemble frameworks the correlation between independent classifiers indicates an accurate classification (Jaffe et al. 2016; Platanios et al. 2017). We assume that a base mask classifying a specific cluster is independent of all other base masks identifying the same cluster in the data, as they are drawn from independently generated unsupervised clustering realizations. Thus independent cluster base masks with $g_{\text{sum}}$ values that are very similar indicate that the cluster realization is accurate. Therefore, the same cluster classifiers will have similar goodness of fit measures; a large change in its value indicates a non-accurate cluster base mask or a physically different cluster group. Clusters with a high number of pixels will have common pixels with the majority of other clusters. This renders very small, but still non-zero, values for the value of $s_I/q_U$ for each comparison between the large cluster and other clusters. This will accumulate value to the masks $g_{\text{sum}}$. Hence there is a bias for large clusters to have a higher ranking values in the ordering of $g_{\text{sum}}$ .

In this section we described the metrics used in the Statistically Combined Ensemble (SCE), which is an unsupervised ensemble framework. Ensemble frameworks have been developed for improving the performance of Artificial Neural Networks (ANNs) and other learning algorithms for decades (Hansen & Salamon 1990; Freund & Schapire 1996; Rokach 2009). Instead of applying a single ANN, an ensemble framework applies multiple ANNs independently on the same input data. It obtains a set of independent classifications for the input data and then combines these classification votes in order to get a joint decision. Each model in the ensemble evaluates the same data and carries out the same task. An artificial neural network desires to reach global optimum, but the end result is highly dependent on the initialisation of weight vectors of the neurons, initial conditions, sampling of the input sample and process parameters of the algorithm. Each network in an ensemble will make errors on different subsets of the input space and thus the ensemble's collective decision is more likely accurate. Thus, an ensemble combines a set of models in order to improve the performance of a single model output. This indicates that the accuracy of an ensemble ANN outperforms the performance of a single ANN (Hansen & Salamon 1990).

# 3 Application to physical datasets

This chapter presents the data-driven problems addressed with clustering algorithms that were discussed in Chapters 1 and 2, and presents the most significant results. Clustering analysis techniques were applied to analyse spatial clustering of observational galaxy data as well as for semantic segmentation of structures from astrophysical plasma simulation images. For both of the case studies, the datasets and their scientific motivation are described first. Then, the selected analysis methods are applied, their performance discussed and the obtained results presented.

## 3.1 Observational galaxy datasets

### 3.1.1 Large-scale structure of galaxy distributions

Cosmic fluctuations and gravitational collapse drive the evolution of objects and the structure formation in the Universe (Sunyaev & Zeldovich 1970). Matter collapses into low-density sheet-like elements; after which matter collapses into long string-like filaments. Then, matter collapses along all directions and creates roughly spherical high-density objects, galaxy clusters. This representation of the previous matter collapse in the Universe is called the Zeldovich pancake picture. The growing gravitational instability in the large scale structure of the Universe is creating the featured contrast in the matter distribution (Peebles 1980).

The cosmic web is a network of all matter in the Universe. It consist of the invisible dark matter and the visible baryonic matter. Galaxies demonstrate the matter distribution of the Universe, driven by the gravitational instability. To understand the evolution of galaxies and the true nature of all matter in the Universe, the model of the large-scale structure of the Universe is constructed. For studying the large scale structure of the Universe, galaxies can be treated as points in 3-dimensional Euclidean space, with Cartesian coordinates of the galaxies being the point locations in space. Other properties of the galaxies are viewed as marks. This has been done for decades by mapping galaxies from large-scale redshift surveys such as the SDSS, 2MASS, 2dFGRS (Martinez & Saar 2002; Davis & Peebles 1983). In the first part of this thesis we will continue to study the structure of the Universe. We discuss the statistical tools used to research the galaxy filament catalogue and two galaxy datasets. We apply these techniques to detect spatial correlation and clustering inside and between these cosmic objects. Namely, we investigate a catalogue of galaxy filaments of the cosmic web, which are detected from the spatial distribution of spectroscopic galaxies in Tempel et al. (2014). Firstly, we analyse the spatial distribution

of galaxies residing in these structures. We underpin correlation between galaxy pairs residing inside galaxy filaments. The detected pattern highlights environmental effects on the formation and evolution of galaxies inside these structures. Secondly, we research an erroneous photometric redshift galaxy catalogue from the SDSS and show their spatial dependence to the detected galaxy filaments. We detect a spatial clustering between galaxy filaments and the independent set of photometric redshift galaxies. The result confirms the modelled structure of the Universe and emphasizes the photometric galaxies possible contribution for the detection of the network.

### 3.1.2 Data and motivation

The datasets analysed for spatial correlation addressed in Paper I and Paper II are the following: i) the Sloan Digital Sky Survey (York et al. 2000) data release 10 of spectroscopic galaxies; ii) the filamentary pattern catalogue detected from the spatial distribution of the previously mentioned spectroscopic galaxies using the method described in Tempel et al. (2014). Spatial locations of objects in both of these catalogues are defined in the 3−dimensional Euclidean space.

Figure 3.1 is a visualization of these two catalogues. We will present results for filaments that have at least one point in the distance range of $100 - 250$ $h^{-1}$Mpc. The filamentary spines in the Bisous dataset (Tempel et al. 2014) have a wide range of possible lengths. This is shown in the paper Tempel et al. (2014) and in Paper I. For the rest of the analysis we are using filamentary spines that are at least $30$ $h^{-1}$Mpc long. This leaves us with 165 filaments, independent of their orientation to the line of sight.

We were looking at galaxies that reside inside these 165 galaxy filaments. A galaxy was defined to belong inside the filament, if it resided at a given distance (0.5 Mpc) from the mathematically detected spine. This gives us 5274 spectroscopic galaxies. All of these galaxies were projected on the filamentary spine that it belonged to.

We determined the preferred positioning of these galaxies along the filamentary spines. This means that we analysed the distances between the red points belonging to an observed purple segment in Figure 3.1. The goal was to detect the preferred pattern of distances between galaxies residing inside filaments. Such a relation sheds light on the environmental processes governing the evolution of galaxies. From the analytics point of view, we were investigating the spatial correlation of pair-wise distances between points in a point pattern. The spatial point pattern is here the galaxies that reside inside the filamentary network. Paper I gives a rigorous description of the catalogues and different filtering methods that were applied on the datasets. Here, we will highlight the main results about the performance of the spatial analytical tool used.

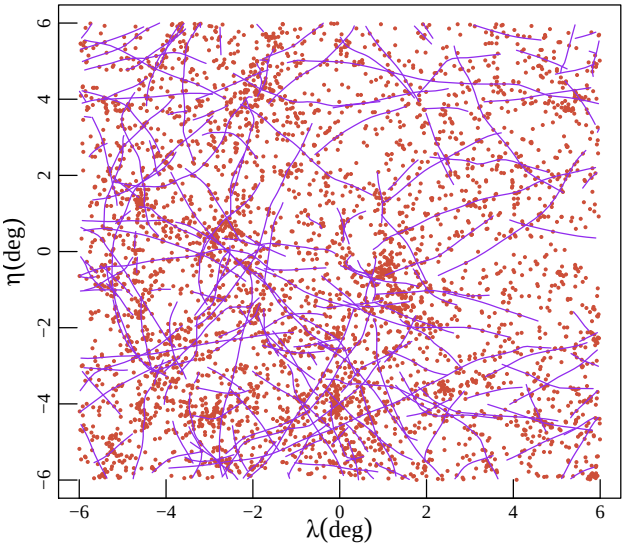



**Figure 3.1**: Visualisation of the datasets in spherical sky-coordinates: filamentary spines (purple spines) and spectroscopic galaxies (red dots). All the drawn objects are located in the distance range of 79 − 277 Mpc.

In Paper III we analysed for spatial clustering the following datasets: i) the filamentary pattern catalogue detected from the spatial distribution of the SDSS (York et al. 2000) data release 12 spectroscopic galaxies with the method described in Tempel et al. (2014); ii) a catalogue of photometric redshift galaxies (Beck et al. 2016). The locations of photometric galaxies are defined in spherical coordinates and the filamentary spines have spatial locations defined in the 3−dimensional Euclidean space.

The distance distribution of photometric galaxies, spectroscopic galaxies and filamentary spines is depicted on Figure 3.2. As expected the distribution of filaments follows strictly that of the spectroscopic galaxies. It is clearly seen that the majority of the photometric galaxies in the catalogue resides in high distances, whilst the spectroscopic galaxies catalogue describes near-by galaxies.

We want to determine whether the photometric galaxies (black background density on Figure 3.3) cluster with the galaxy filaments (purple segments on Figure 3.3). The interest is to determine whether the photometric galaxies would possess information about the filamentary network, which are detected from the spatial distribution of spectroscopic galaxies. Paper III gives a rigorous description for the catalogues, filtering methods applied and feature extraction done for the catalogues.

The redshifts describing galaxy location of photometric galaxies are rather un-

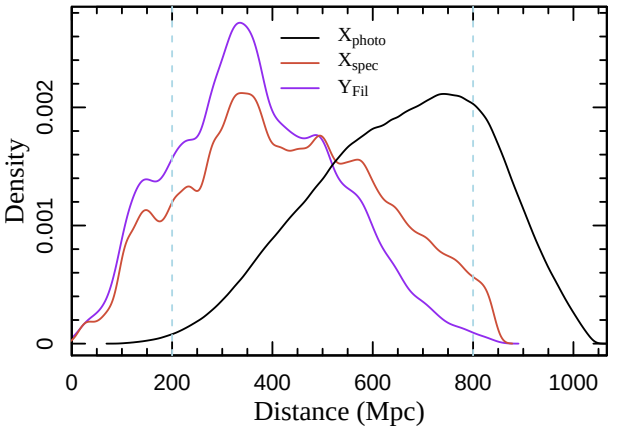



**Figure 3.2**: Visualization of the distribution of distances for spectroscopic redshift galaxies (red line), the detected filamentary spines (purple line) and the photometric redshift catalogue (black line). Vertical blue dashed lines indicate the distance range of the analysis.

certain, error $\delta z_{\text{photo}}$ reaching up to 0.05. This results in the absence of an accurate 3−dimensional coordinate system for the galaxy positions. Corollary, they were not used in the initial modelling of the filamentary network in Tempel et al. (2014). Thus it is of great interest to determine whether this galaxy catalogue presents new information about the structures. The analysis determines whether the galaxy filaments detected from the spatial location of spectroscopic galaxies is confirmed to be located in the same areas of space by the photometric redshift galaxies. In addition, these photometric galaxies estimate the properties of galaxy filaments, such as their most preferred physical radii and types of galaxies most likely residing in filamentary networks.

A thorough clustering analysis is done in Paper III addressing all the aforementioned questions. The galaxy filaments were filtered dependent on their orientation in the line of sight and the photometric galaxies were sampled into multiple distance segments. In addition, the galaxy filaments were represented by the spectroscopic galaxies residing in them, and not only by a configuration of segments building up the filament spine. For the 3−dimensional representation of physical distances between the objects, a 3−dimensional smallest-distance distribution function was derived. In the 3−dimensional case, photometric galaxies were represented by lines of sight going through their spherical coordinates, such that the galaxy might be located at any point along the line following through its position on the sphere. The 3−dimensional smallest-distance distribution estimated physical properties of the filaments, such as

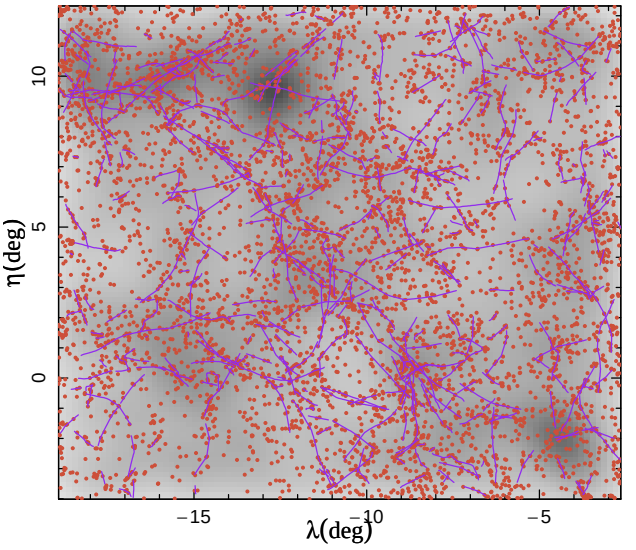



**Figure 3.3**: Visualization of spectroscopic redshift galaxy (red points), the detected filamentary spines (purple segments) and the photometric redshift catalogue (black background density). Darker color denotes higher density of photometric galaxies.

its thickness and bias towards more massive galaxies residing in them.

In this thesis, only the main results of the spatial clustering analysis for the photometric galaxies dataset and filamentary network catalogue is presented. The subsamples of the original datasets are described in the following:

- The photometric galaxies lack precise redshift estimations, which renders the presented spatial clustering analysis to spherical coordinates on a region of a sphere $\mathbb{S}^2$. The galaxies in the dataset of photometric galaxies are represented by their radian latitude $\eta$ and longitude $\lambda$. We will use the photometric galaxies residing in the distance gap $200 - 800$ Mpc, which concludes in 2 198 702 photometric galaxies $X_{\text{photo}_{\text{all}}}$. These photometric galaxies are viewed as a configuration of points $X_{\text{photo}_{\text{all}}}$ in a region of a sphere $\mathbb{S}^2$.

- The distance gap of $200 - 800$ Mpc leaves us with a set of 38 702 filamentary spines $Y_{\text{fil}_{all}}$. Each filament in $\mathbb{R}^3$ was mapped to the spherical coordinates latitude $\eta$ and longitude $\lambda$ on $\mathbb{S}^2$. They form a random set of objects $Y_{\text{fil}_{all}}$ in a region of a sphere a $\mathbb{S}^2$. The filaments were sampled according to their orientation towards the line of sight $\bar{\alpha}(Y)$, a metric defined in Paper III. Value $\bar{\alpha}(Y)$ represents the average angle that the spine $Y$ has with respect to the line

of sight. This was to determine whether the orientation of the filaments has influence to the clustering.

This means that we will study spatial clustering between a point pattern (photometric galaxies) and a random set of objects (spines of filaments) on a region of a sphere $\mathbb{S}^2$. The bivariate $J-$function from the theory of spatial point patterns will be used to address this question.

### 3.1.3 Analysis and results

**The pair correlation function**

Figure 3.4 illustrates the pair correlation function for the pair-wise distances of galaxies residing in filamentary spines. The depicted result is also seen in Paper I and Paper II. The high peak for the smallest-distance range highlights the abundance of galaxy

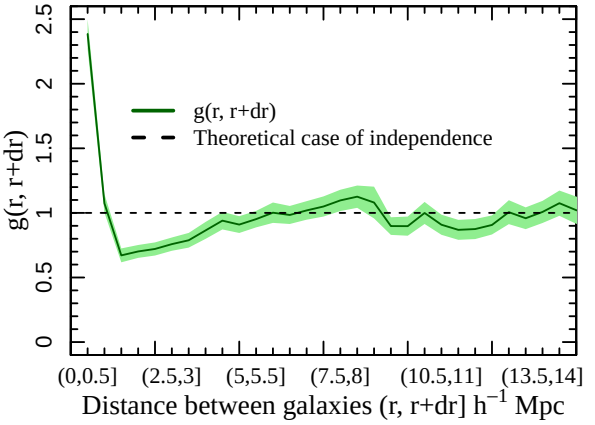



**Figure 3.4**: The pair correlation function for galaxies inside filamentary spines (dark green line), the Jackknife confidence interval (light green) and the theoretical case of independence (dashed black line).

groups in the dataset, as galaxies in galaxy groups have small pair-wise distances. The dip in the value of the estimator around $1.5 - 2.0\,\mathrm{h}^{-1}$Mpc indicates a lack of galaxies around the galaxy groups. The most interesting and a statistically significant peak in the values of the pair correlation function is around $7.5 - 8.0\,\mathrm{h}^{-1}$Mpc. This peak indicates that this distance range exists with higher probability in the catalogue of galaxy pair-wise distances in comparison to the catalogue of randomly distributed points pair-wise distances. The physical explanation for this is not obvious, but one can speculate that it is connected to galaxy evolution and mergers inside filaments.

Peaks follow periodically, which is due to a periodical location pattern of galaxies along long filaments.

The pair correlation function that studies distances between galaxies along filamentary spines revealed clear statistically significant patterns. The estimator is easy to interpret and compares the catalogue directly to complete spatial randomness. Our analysis showed that galaxies tend to locate along the filamentary spines like pearls on a necklace.

**The bivariate $J-$function**

As mentioned in Section 3.1.2 an elaborate analysis of clustering on a region of a sphere $\mathbb{S}^2$ and in 3−dimensional space between the different subsets of the catalogues is done in Paper III. In the following we will present the main result obtained with the bivariate $J-$function. This study clearly highlighted the potential of the bivariate $J-$function for future astronomical spatial clustering analysis.

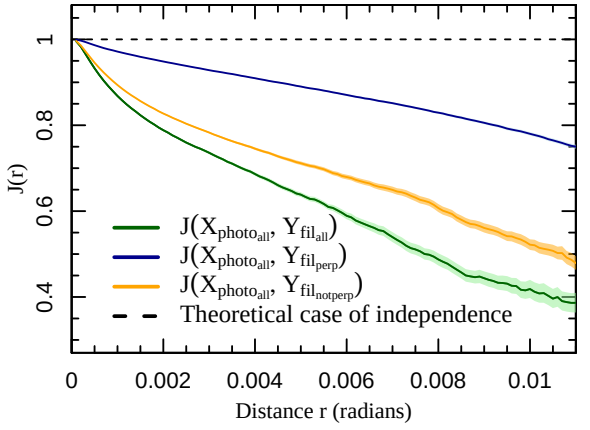



**Figure 3.5**: Results of the bivariate $J-$functions $J(r)_{X_{\text{photo}_{\text{all}}},Y_{\text{fil}_{\text{all}}}}$ (green), $J(r)_{X_{\text{photo}_{\text{all}}},X_{\text{fil}_{\text{perp}}}}$ (blue), and $J(r)_{X_{\text{photo}_{\text{all}}},X_{\text{fil}_{\text{notperp}}}}$ (orange) in comparison with the theoretical reference case representing independence between the studied sets (black dashed line). The radian distance of 0.002 corresponds to 0.4 Mpc at the distance of 200 Mpc and 1.6 Mpc at the distance of 800 Mpc.

The bivariate $J-$function was calculated between the catalogue of photometric galaxies $X_{\text{photo}_{\text{all}}}$ residing 200 – 800 Mpc from the observer and all the following sets of filamentary spines in the distance range of 200 – 800 Mpc:

- all filaments $Y_{\text{fil}_{all}}$,

- non-perpendicular filaments in relation to the lines of sight with $0° \leq \overline{\alpha} < 78.22°$, which concluded in 30 639 filaments $Y_{\mathrm{fil}_{\mathrm{notperp}}}$,
- perpendicular filaments in relation to the lines of sight with $78.22° \leq \overline{\alpha} \leq 90°$, which concluded in 8 063 filaments $Y_{\mathrm{fil}_{\mathrm{perp}}}$.

Figure 3.5 shows the bivariate $J-$functions for photometric galaxies and all filamentary spines in the viewed region, and the bivariate $J-$function for the same photometric galaxies and two subsets of the spines, defined as perpendicular to the line of sight and defined as non-perpendicular to the line of sight. Applying the bivariate $J-$function to study the spatial clustering between a set of filamentary spines and photometric galaxies catalogues provided us with clear clustering signals. In Fig. 3.5 the decreasing $J(r)_{X_{\mathrm{photo}_{\mathrm{all}}},Y_{\mathrm{fil}_{\mathrm{all}}}}$ values below the theoretical reference case represent a positive association between the photometric galaxies and all filamentary spines. Positive association is clearly shown by the lines of $J(r)_{X_{\mathrm{photo}},X_{\mathrm{fil}_{\mathrm{perp}}}}$ and $J(r)_{X_{\mathrm{photo}},X_{\mathrm{fil}_{\mathrm{notperp}}}}$ as well. This indicates that the photomeric galaxies are positively associated (clustered) with filaments independent of the orientations $\cos(\bar{\alpha}(Y))$ (defined in Paper III). The faster decline in the $J-$function as a function of $r$ for photometric galaxies $X_{\mathrm{photo}_{\mathrm{all}}}$ and all filamentary spines $Y_{\mathrm{fil}_{\mathrm{all}}}$ indicates a stronger positive association between these spatial patterns.

The bivariate $J-$function clustering analysis verified that the information hidden in photometric redshift galaxies can contribute to the filamentary network detection. Even with uncertain redshift estimations and projection effects, the study of spatial spacings gave insight to the importance of the galaxy data in structure modelling of the Universe. We showed that large-scale photometric redshift estimates can contribute greatly to the modelling of the filamentary network. However, until now the redshift estimates are too uncertain. In upcoming surveys, such as the Javalambre Physics of the Accelerating Universe Astrophysical Survey (J-PAS) (Benitez et al. 2014; Bonoli et al. 2020), these galaxies will become even more important in the modelling of the cosmic web structure. Additionally, the clear clustering signal of the detected filamentary network and a big new dataset of galaxies, highlights the importance of detecting the filamentary structures. Majority of galaxies in the Universe reside in these structures. Seemingly, in order to understand their evolutionary process it is of utmost importance to determine and understand these environments.

## 3.2 Plasma simulation dataset

### 3.2.1 Kinetic turbulent collisionless plasma

Kinetic, turbulent, collisionless plasma can be used to model physical phenomena observed in e.g., stars, black-hole magnetospheres, supernova remnants and galaxy

clusters. They are applicable also in Earth-bound settings, such as in modelling plasma in fusion reactors. For a thorough description of the underlying physics of turbulent flow structures in plasmas we refer to the works of Zhdankin et al. (2017), Comisso & Sironi (2018), Nättilä (2019) and Uritsky et al. (2010). This section starts by giving a brief overview of the physics of turbulent collisionless astrophysical plasma simulations using Nättilä (2019) and Paper IV.

Nättilä (2019) simulates such a kinetic turbulence in a suddenly stirred magnetically-dominated plasma. The geometrical properties of the merging round islands and thin elongated physical structures that start to evolve in the spatio-temporal plasma are of great interest. The thin elongated structures are thought to be the driving force for particle decoupling from the plasma pool and are physically not fully understood. Their detection is still an open question and a lot of new physics is yet to be learned from their statistical analysis.

To model a turbulent magnetically dominated plasma, consider a uniform plasma of charged particles with density $\rho$ governed by an initial guide field $\vec{B}_0$. The guide field $\vec{B}_0$ is typically aligned to some externally defined direction or a large-scale field, that fixes the plasma motions to be transverse to this direction. This renders the problem basically $2-$dimensional as only the planar directions evolve. By stirring the plasma at some large scale, round eddies start to form and the plasma coagulates into solitary magnetic "islands". These are round magnetic tubes, which are oriented mainly along the initial guide field $\vec{B}_0$. If these islands merge, a strong current $\vec{j} \parallel \vec{B}_0$ is formed to balance the suddenly flipping transverse magnetic field. These current sheets will be torn apart by a tearing instability, which leads to a reconnection of the magnetic field. Energy released in this process is absorbed into excitations of high-energy particles. The magnetic energy is then transferred into kinetic energy of the particles, which then decouples the aforementioned high-energy particles from the thermal plasma pool. This decoupling then provides an energy dissipation mechanism that is powered by the current sheets. This underlines the importance of being able to identify these thin sheet-like structures: Detecting them would enable us to study the dissipation processes in situ.

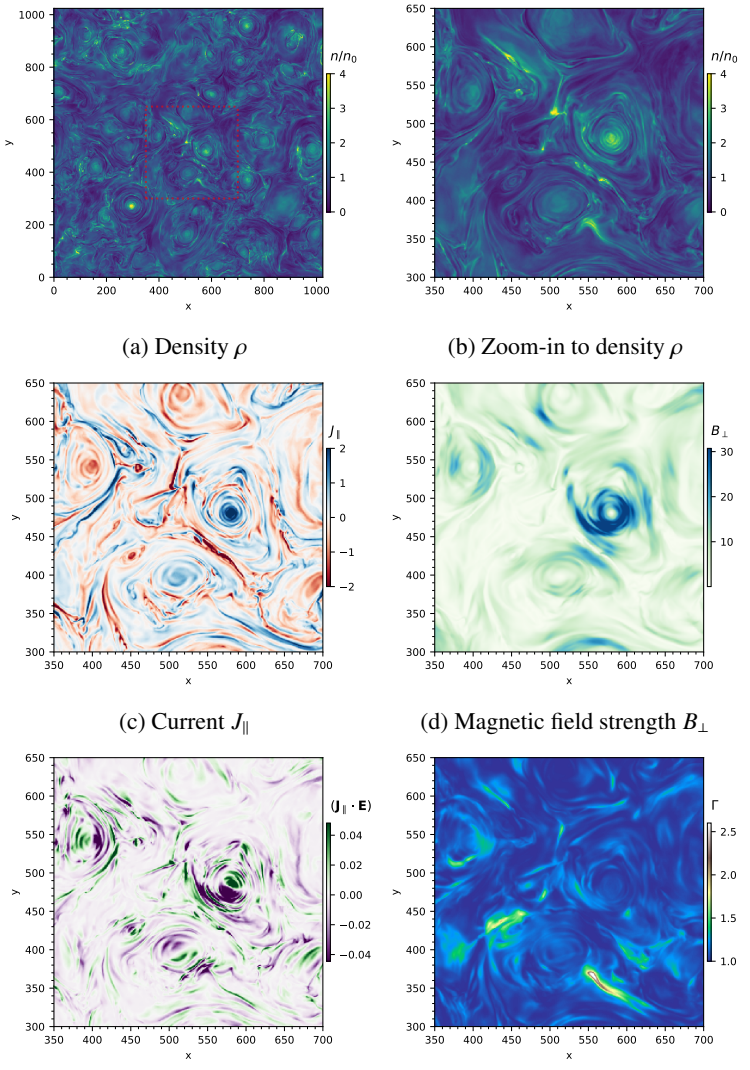
In this thesis, we will apply a well-known unsupervised clustering algorithm and develop an ensemble framework to address the stochastic nature of the artificial neural networks. With these tools we are able to automatically segment the thin elongated structures from the plasma pool with pixel-level accuracy. These dissected physical structures are further studied for their geometrical and physical properties in papers currently being carried out.

### 3.2.2 Data and motivation

The dataset analysed was a spatio-temporal dataset of magnetically-dominated collisionless plasma images obtained from Nättilä (2019) kinetic particle-in-cell simulation. The simulation data can be presented in the form of consecutive images, describing the particle-level evolutionary progress of the original plasma. Each pixel contains physical observable values that change in each snapshot during the course of the simulation.

Figure 3.6 presents a snapshot of the simulation data. The figure shows five physical features grasped by the computer simulation. The Figure 3.6a shows the whole simulation box, it consists of a little over 1 million pixels. The depicted feature is the plasma density $\rho$ at simulation snapshot 5000. Rest of the images are zoom-in views to the simulation box. Figure 3.6b shows the plasma density $\rho$ in the zoom-in view. Circular islands and thin lines are distinctly visible in this view, with a higher number of particles residing in them. Figure 3.6c shows the direction and strength of the current $J_{\parallel}$ (parallel to the initial guide field $\vec{B}_0$) in the zoom-in region of the plasma snapshot. We see that the same islands and the thin lines have high values of current shooting out or in to the plane of the image. Figure 3.6d shows the strength of the perpendicular (to initial guide field $\vec{B}_0$) magnetic field strength $B_{\perp}$. The large islands are distinguishable from the pool of plasma by the high value of the perpendicular magnetic field strength. Figure 3.6e is of the work done by the electric field ($J_{\parallel} \cdot E$) in the zoom-in view and Figure 3.6f is the $\Gamma$, which describes the Lorentz factor of each particle. The circular islands are clearly visible in the ($J_{\parallel} \cdot E$) view and the thin lines have larger Lorentz factors. All the features have multiple self-similar structures that are clearly visible on each of the images on Figure 3.6. These are the observed circular structures referred to as eddies that coincide with a high signal in $\rho$, $B_{\perp}$ and in $J_{\parallel}$. Another prominent feature are the thin stripes referred to as current sheets that correspond to a maximum in $J_{\parallel}$, $\Gamma$ and a minimum in $B_{\perp}$. In the simulation images the most prominent current sheets are visible in between eddies, which are in the process of a merger. The turbulence cascade creates structures of a wide range of sizes, the biggest are clearly easily visible on all of the images in Figure 3.6.

The aforementioned physical structures are clearly visible in the astrophysical plasma simulation data, but they are difficult to detect automatically. One reason being that the structures in the plasma have a chain of varying sizes, another being that the physics creating the shapes is not thoroughly understood. In the long run, we are interested in obtaining precise estimations for the sizes of these magnetic islands and the elongated current sheets. Their geometrical properties are of great interest in the astrophysics domain. The reliability of pixel-by-pixel dissection of structures from a plasma pool is of even greater importance, as the results we obtain will be used to understand the underlying physics in the shapes.




(a) Density $\rho$

(b) Zoom-in to density $\rho$

(c) Current $J_\parallel$

(d) Magnetic field strength $B_\perp$

(e) Work done by electric field $(J_\parallel \cdot E)$

(f) Particle's Lorentz factor $\Gamma$

**Figure 3.6**: Visualization of the turbulent plasma simulation at snapshot 5000. Each pixel has five different features that we visualize here. The top left panel shows the full image domain whereas the rest of panels show a close-up region of the image (marked with red-dashed line in the top left panel). Multiple types of structures manifest in the different features views of the plasma, most prominent structures being the circular islands and thin stripes.

Our main goal in Paper IV was to obtain robust region-of-interests (ROI) for the aforementioned structures. As a data analysis concept this meant finding similar pixels in the physical feature space. Similar objects can be detected from the feature space with clustering algorithms, which group together objects that are more similar to each other than the ones laying outside of the group. In addition, we needed to take into account the neighborhood of the pixels, as we are interested in geometric objects on an image. Thus, the clustering algorithm should be sensitive to the topological space of the input image.

One simulation snapshot image consists of a little over 1 million pixels. We perform the clustering by simultaneously using 8 consecutive snapshots from the simulations, with equal time steps between the samplings. This brings the total number of data points to over 8 million pixels, each with a 3−dimensional data vector $X_k = (J_\parallel, B_\perp, (J_\parallel \cdot E))$. These characteristics were chosen from a physical knowledge of the current sheets as they describe most of the variability in the catalogue. The analysis was carried out also on a higher and lower count of consecutive images. The number of 8 images, separated by a fixed time step, was chosen by trial and error. The images at the chosen time steps exhibited the physical structures that were most important from the astrophysical viewpoint. In the beginning of the simulation (earlier time steps) the artificially inserted set in turbulence is prominent. The cascade of the plasma structures has not yet created the cascade of self-similar structures. When the simulation evolves further, it evolves into a more stable state and the structures are not as prominent.

### 3.2.3 Analysis and results

**Self-Organizing Map**

We applied an artificial neural network, the Self-Organizing Map (Section 2.2), on the image data from the plasma simulations. We trained a 2−dimensional grid of neurons to represent the high-dimensional input data space. The count of neurons on the neuron layer was significantly lower than the count of input vectors in the input image data. Each neuron was assigned with a weight vector, which are initialized randomly from the input sample space. During the learning process neurons will learn from the input sample space and adjust their weight vector according to the learning rule. The algorithm is stochastic and the neural map changes with each iteration step, as it learns to represent the distribution of the input sample space with a 2−dimensional representation. The algorithm converges to a result dependent on its initial process parameters, internal functions, random initialization and the random choice of input vectors. The neighbourhood function $h_{ci}(t) \to 0$ as $t \to \infty$ in Equation (2.2), which grants that the map will converge and no significant changes are created with further iterations. Kohonen (2001) emphasizes the importance of iteration steps for the

convergence of the model. Mathematically two and higher dimensional SOM neural maps are not guaranteed to converge to a global optimum (Erwin et al. 1992). It is possible to test, whether further iterations change the map significantly, but on the other hand, the map might have converged only into a local optimum and not into the desired global optimum. This is a problem known for artificial neural networks at large and are discussed in the concept of SOMs in many previous works, for example in De Bodt et al. (2007) and Lee & Verleysen (2002). Additionally, training the algorithm for too many iterations might lead to over-training. In this case, the neural network becomes too specific to the input data sample and performs poorly on new data.

We encountered the aforementioned obstacles in the clustering analysis of plasma simulation images: we found it difficult to converge to a global optimum and saw the algorithms volatility to the choice of process parameter values.

We applied the SOM algorithm to 8 consecutive simulation image snapshots and we sampled the parameter value space for the Self-Organizing Map algorithm. The parameter combinations sampled for our particular analysis are:

- Rectangular shaped neuron map $(m, n)$, with dimensions $(15, 10)$.
- The learning rate $0 < \alpha(t) < 1$ values are chosen to be $\{0.6, 0.7, 0.8\}$.
- Number of total iterations (i.e., the training steps) is chosen to be 10 000, 20 000, 30 000, 40 000, or 50 000.

All the possible aforementioned parameter combinations of the SOM algorithm are applied on the studied astrophysical plasma simulation images. This gives 15 independent SOM cluster ID results for the same images. Figure 3.7 visualizes the results from four representative SOM clustering algorithm outcomes. The retrieved cluster ID for each pixel is projected to the original image view. Each of these maps were evaluated on the same 8 million pixels of data with differing process parameter values. The images of cluster maps on Figure 3.7 can be directly compared to Figure 3.6, where the same snapshot of the original training image is illustrated. The clusters detected by the SOM algorithm on Figure 3.7 correspond to distinct regions in the original images: circular island regions, thin stripes, and large background areas that are also visible in Figure 3.6.

Importantly, the resulting clusters differ for each SOM outcome, as the output of the algorithm is dependent on the randomized nature of the initial neural map, sampled input data, and the process parameters. Especially the geometric sizes of resulting segmented cluster regions differ significantly between these different realizations. We applied the SOM with multiple combinations of process parameters on the input image, but found it hard to converge and retrieve a cluster map that did

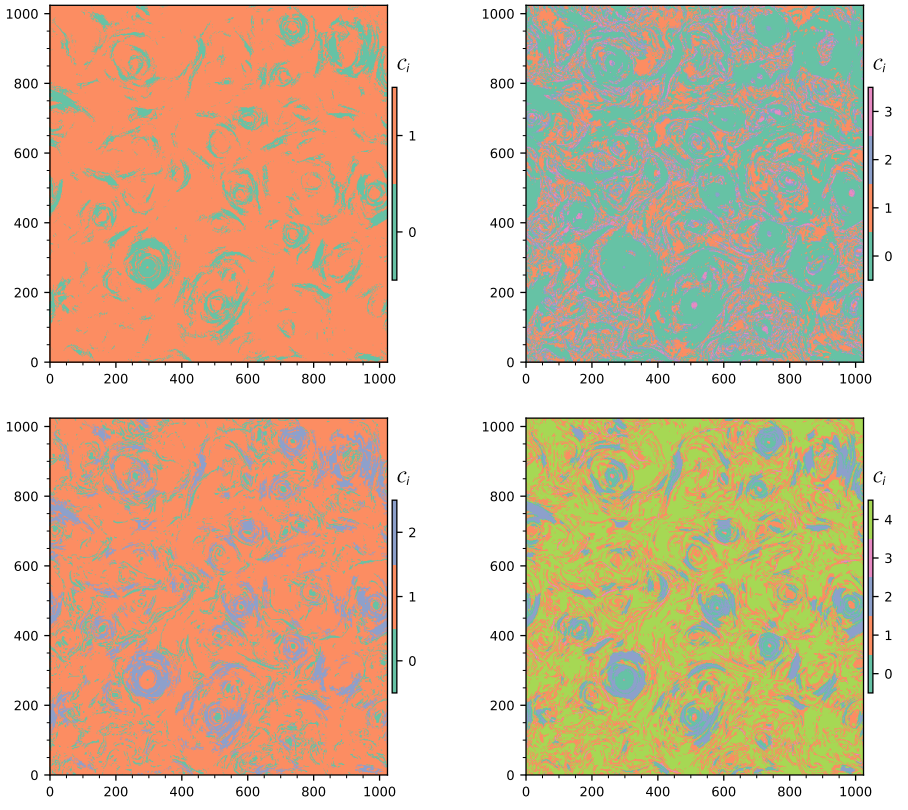



**Figure 3.7**: Visualization of four different SOM evaluations projected back to the original image view. These four SOM clustering results are shown for the simulation snapshot 5000 depicted in Figure 3.6. Each of the SOM runs group the initial data pixels into $n$ clusters, $C_i$. We color the pixel based on the cluster it is associated with in order to visualize the physical structures that the data clusters correspond to.

not change significantly between different evaluations. This indicates that analysing the dissected structures from one cluster map realization only might be misleading. This is unfortunate especially as the geometrical properties of these structures are the prominent information we want to obtain from the segmentation analysis. The preliminary results are evaluated by the physical interpretation and visual inspection of the detected structures. As there exists no ground truth detection of the physical structures from the corresponding simulation images, the quality of the results is not so easily quantifiable. In the supervised domain, the Mean Intersection over Union (MIoU) is often used to estimate the performance of the algorithm. In the following section we will compare and order the results of independent SOM algorithm realizations with the metrics proposed in Chapter 2.3. These metrics, in their essence, highlight the SOM output which has learned the most from the initial input sample.

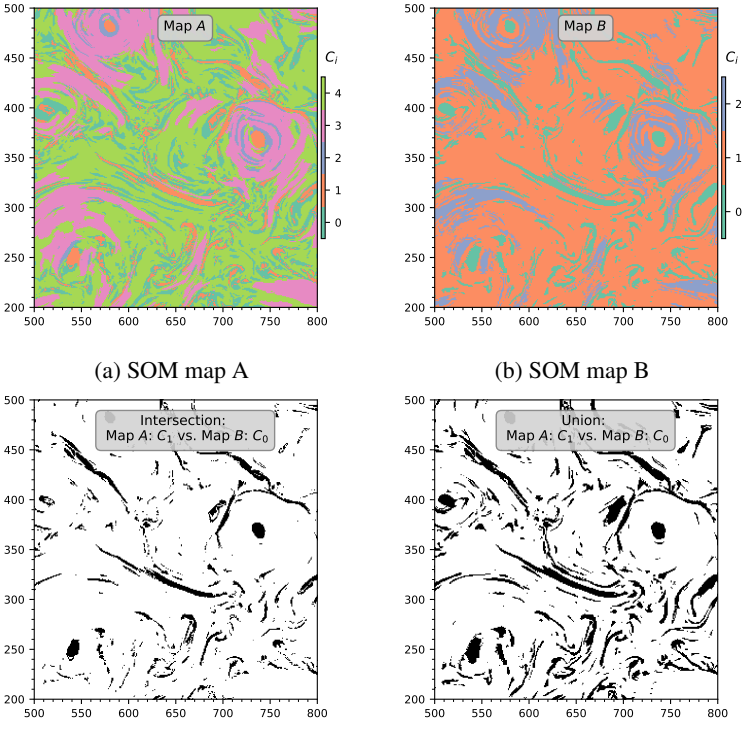
In addition to the SOM algorithm the k-means and k-nearest neighbours algorithm (Cover & Hart 1967) were applied. The SOM algorithm performed better as it was able to maintain the topological information present in the input sample space and it was able to detect fine structures from the input images. The SOM compresses the information stored in the input data, but doing so it tries to preserve the topology of the input space. The goal of image segmentation of the astrophysical plasma images was to dissect continuous physical structures. The topology of the input data is thus important to be preserved. In future work with the data deep neural network algorithms will be applied. The SOM provided a simple enough gateway to the usage of artificial neural networks.

**Statistically-Combined Ensemble of cluster masks**

The results obtained with the SOM algorithm together with the observed shortcomings motivated the Statistically-Combined Ensemble (SCE) framework, where we stack independent cluster maps of the structures. The SCE framework is mathematically described in Section 2.3

The SCE method is a multi-map comparison framework that aims to alleviate the map-to-map variations in structure borders on the image by statistically combining many independent SOM cluster evaluations. In our case we stack a set of 62 SOM cluster masks, which were obtained from 15 independent SOM runs with different process parameter combinations. The similarity measures and goodness-of-fit metrics described in Section 2.3 were used to find stable regions of interest (ROI) for the physical structures in the image.

As an example, Figure 3.8a and Figure 3.8b visualize two cluster maps of two independent SOM clustering results labeled as map $A$ and map $B$. Map $A$ has detected 5 clusters and map $B$ 3 clusters. Thus, the corresponding mask sets are $\mathcal{M}^A = \{M_0^A, M_1^A, M_2^A, M_3^A, M_4^A\}$ and $\mathcal{M}^B = \{M_0^B, M_1^B, M_2^B\}$. We will calculate and vi-




(a) SOM map A

(b) SOM map B

(c) Intersection matrix of $M_1^A$ and $M_0^B$

(d) Union matrix of $M_1^A$ and $M_0^B$

**Figure 3.8**: Visualization of the map-to-map operations performed between two SOM cluster map realizations. Top panels show a comparison of two SOM runs (focusing on a small region of the complete image) for which the pixel group information is projected back to the original image view. The SOM map A detected 5 clusters and the SOM map B detected 3 clusters. Lower panels show the intersection and union matrix for $C_1$ cluster mask of the SOM map A and $C_0$ cluster mask of the SOM map B. The clusters in both maps have captured the thin stripe structures but the exact cluster boundary locations vary slightly.

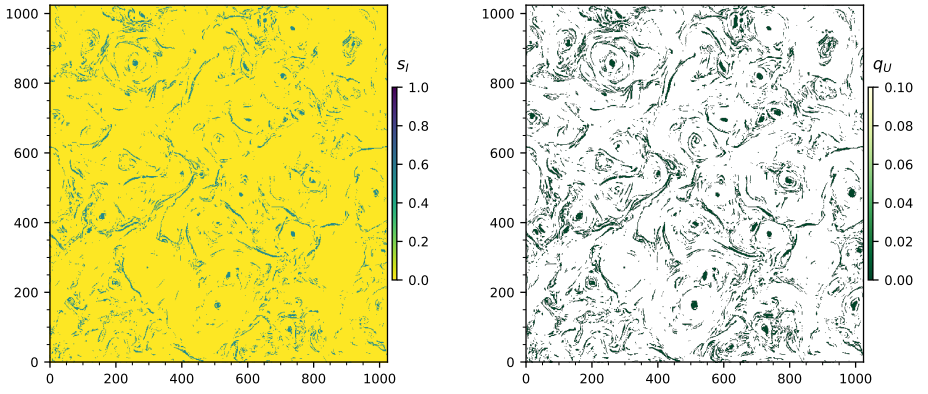



**Figure 3.9**: Resulting signal strength $s_I$ (left panel) and quality $q_U$ (right panel) values computed between the two clusters of map $A$ and map $B$ (visualized in Figure 3.8). Signal strength $s_I$ is a measure of the area in both masks contributing to describing the same structure. Quality value $q_U$, on the other hand, is a measure of the residual area between the union and intersection of the two masks. A small residual between the masks ($q_U \approx 0$) and similar area ($s_I \approx 1$) together correspond to a well-matching mask combination.

sualize the intersection matrix $I$ (Equation (2.5)) and union matrix $U$ (Equation (2.4)) for the cluster mask $M_1^A$ from map $A$ and cluster mask $M_0^B$ from map $B$. These matrices are depicted as images on Figure 3.8c and Figure 3.8d. The pixels colored black correspond to the value 1 and pixels colored white correspond to 0. The $I$ matrix highlights the pixels assigned to $M_1^A$ (cluster $C_1$ mask of map A) and $M_0^B$ (cluster $C_0$ mask of map B). The $U$ matrix activates all pixels belonging to both masks, $M_1^A$ and $M_0^B$. For the maps $A$ and map $B$ the $I$ and $U$ matrices are similarly shaped and positioned structures, suggesting that the maps have detected the same structure from the input image.

The signal strength $s_I$ (Equation 2.7) for base mask $M_1^A$ and a mask $M_0^B$ is visualized on the left panel of Figure 3.9. The metric compares the intersection and union of map $A$ cluster $C_1$ mask and map $B$ cluster $C_0$ mask, shown on Figure 3.8c and Figure 3.8d. It describes the fraction of overlap between the two cluster masks; value of the quantity is, naturally, the highest on the location where the two masks overlap.

The right panel of Figure 3.9 visualizes the quality of the union metric $q_U$, which is defined by Equation (2.8). The measure quantifies how well the two cluster masks $M_1^A$ and $M_0^B$ align on top of each other. In case of a perfect alignment, the value of the quantity approaches 0; this would correspond to identical masks. In Figure 3.9 pixels with values close to 0 indicate a good fit between the two masks.

For perfectly overlapping two cluster masks, meaning two independent SOM evaluations having detected equal amounts of pixels in the same positions on the image, we would see $s_I \to 1$ and $q_U \to 0$ for their mask comparisons. These cluster comparison metrics are negatively correlated, first giving an estimate to the strength of the intersection of the masks and the other for the size of the symmetric difference of the masks. These metrics can be used separately to gain perfectly overlapping cluster masks and cluster masks detecting completely different structures ($q_U \to 1$).

In our study we wanted to establish robust borders for structures detected on the image, such that a detailed analysis of their geometry could be performed. Thus we stacked multiple independent cluster masks on top of a randomly chosen base cluster mask. This was done until all cluster masks from each SOM evaluation was chosen as the base mask. The stacking of multiple cluster masks was done with the use of the goodness-of-fit matrix $G_{sum}$ (Equation (2.11)), which added all the quotients of the $s_I$ and $q_U$ applied to the union of the two compared masks. This summing accumulates value to pixels on the image with each comparison, such that the cluster mask that has detected a similarly sized structure in the same locations as the base mask will contribute significantly to the total sum. On the contrary, a cluster mask detecting a different structure contributes a negligible amount to the total. This will highlight most stable pixels belonging to that structure.

All 62 cluster masks detected by 15 independent SOM algorithms will result the $G_{sum}$ matrix, which will give a value to each pixel on the image such that more robust borders for the structure can be obtained. These 62 cluster masks can be ordered according to their scalar sum of the $g_{sum}$ (Equation (2.12)). The base masks detecting a cluster, which has been detected by many other independent algorithm runs will have a higher $g_{sum}$ scalar value. This is due to the fact that it will have higher quotients of $s_I$ and $q_U$ for mask comparisons, which accumulate to the total sum. This means that the scalar $g_{sum}$ can be used to detect the base mask detecting the most stable cluster of structures on the image.

Figure 3.10 shows the total integrated scalars of goodness of fit, $g_{\text{sum}}$ (Equation (2.12)) for all the masks detected by the 15 independent SOM evaluations. The clusters have been visually inspected and 3 major physical structures are recovered: background pixels, islands (circular magnetic tubes), and current sheets (thin stripes). The aforementioned structure elements in the plasma were visible in the original plasma simulation images on Figure 3.6.

The background cluster (red points in Figure 3.10) is detected the best among all the independent SOM algorithm evaluations, since the sum of the stacked $s_I/q_U$ values are the highest and the sum is most similar with other independent classifiers for the same cluster. The second best cluster detected is that of the islands, (dark-blue diamonds in Figure 3.10). The cluster masks of current sheets are the third best de-

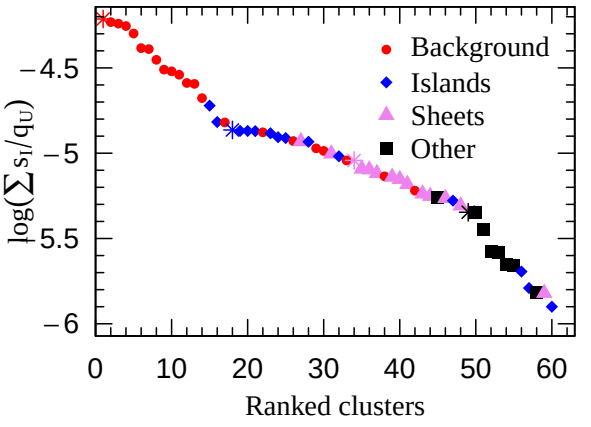
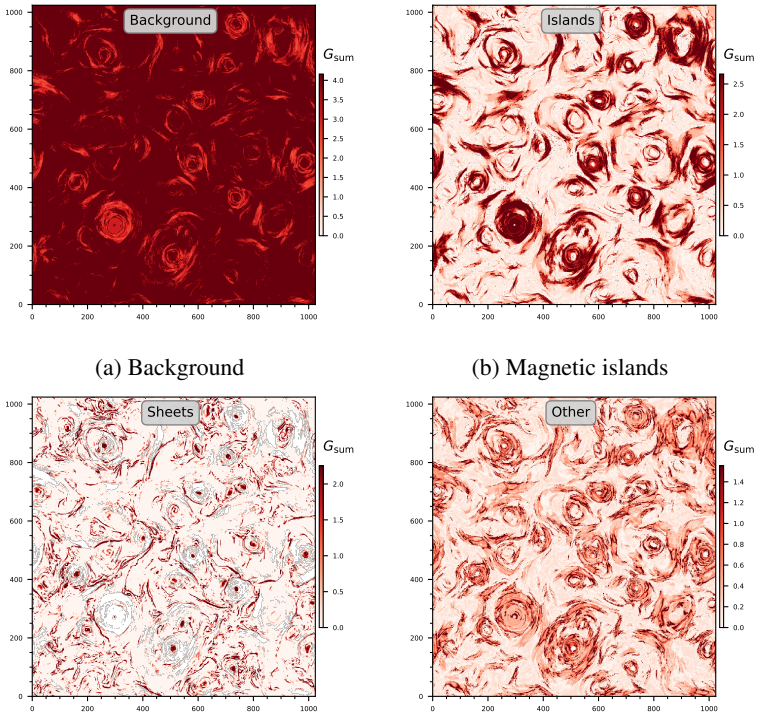



**Figure 3.10**: Total scalar values of the goodness-of-fit, $\log_{10} \sum_{\mathcal{M} \setminus M^b} (\frac{s_I}{q_U})$, (see Equation 2.12) for all the 15 independent SOM runs in the analysis. In total, the collection of maps has 62 clusters. The clusters are shown in descending order. Each result was visually inspected and labeled into three different empirical categories: background pixels (red), circular islands (blue), thin stripes (violet), and other unclassified shapes (black). The cluster mask from each physical cluster set with the highest value and best accuracy is denoted with a star symbol. This manual classification is seen to correlate fairly well with the goodness-of-fit of the specific cluster.

tected structures (violet triangles in Figure 3.10). This visual classification is seen to correlate with the corresponding goodness-of-fit value. Cluster realizations between different SOM evaluations can be grouped together, since base masks detecting a similar structure are expected to have a similar $g_{\text{sum}}$. Additionally, large changes in the value of $g_{\text{sum}}$ match well with the change of the physical meaning of the clusters or indicate that the base mask classifying the observed cluster is non-accurate.

Figure 3.11 represents the stacked cluster maps of the four distinct structures detected on the plasma simulation. These are the four base masks, which corresponded to the highest $g_{sum}$ value in their structure set and are most similar to independent cluster masks detecting the same physical structure (clusters denoted with stars on Figure 3.10). Highly correlated base masks detecting same physical structures indicate a high accuracy for detecting the structure. Each pixel in each of these image views has obtained a goodness-of-fit value. This corresponds to this pixels stability belonging to this cluster. If the pixel was assigned to a cluster mask similar to the base cluster mask in multiple comparisons, then the value acquired on this image is high. The value will be low, if the pixel was mostly assigned to a cluster mask very different from the base cluster masks.




(a) Background

(b) Magnetic islands

(c) Current sheets

(d) Mixture of structures

**Figure 3.11**: Goodness-of-fit matrices ($G_{\mathrm{sum}}$) of the four best detected stacked cluster maps (with $\log_{10}$-scale colors). Each panel shows the highest ranking cluster from Figure 3.10, ordered based on their integrated goodness-of-fit quantity. The three empirically derived categories are clearly visible in the results: background pixels (Figure 3.11a), islands (Figure 3.11b), current sheets (Figure 3.11c). Figure 3.11d is seen to be a mixture of many clusters (mainly circular islands and thin stripes).

Figure 3.11d is correctly ranked as only the fourth best category and could therefore be safely discarded from any further analysis. For the first three images in Figure 3.11 it is straightforward to define a quantitative cutoff level that enables to define structure contours. Since the resulting contour boundaries are averaged over many SOM realizations their boundaries are more robust and hence any geometrical analysis of the resulting shapes is more reliable. A threshold can be applied to the obtained value of $G_{sum}$ to estimate geometrical metrics for the current sheets on the image. The current sheets are detected from the plasma simulation image – the result is clearly visible on Figure 3.11c. The dissected structures are robust and continuous, which can be further analysed for their geometrical properties. This work is being done in an upcoming paper.

By statistically combining cluster evaluations of unsupervised clustering algorithms robust boundaries for structure clusters were obtained. The framework produced cluster information for pixels on an image, accurately detecting the most interesting structures in the plasma, the current sheets. As has been emphasized before, the physical nature of the current sheets is not fully understood. In addition the detection of current sheets from a spatio-temporal dataset of turbulent plasma is an open question. Thus there exists no ground truth information to perform validation for the cluster outputs. The obtained ROI for the structures are visually inspected to follow the expected areas and shapes of the current sheets. Further study of the physical properties of the segmented structures will highlight physical reliability of the results. Overall, this work has been an exploratory step at automating the detection of current sheets from simulation images. The following projects include the physical and statistical analysis of the detected ROI's and further development of the detection process.

# 4 Conclusions

The aim of this thesis was to apply and develop clustering algorithms for astrophysical datasets, in order to understand and also reveal new hidden physics in the data. The largest studied structure in astrophysics is the cosmic web, consisting of hundreds of billions of galaxies. The properties, evolution and mergers of galaxies inside the dark-matter-dominated cosmic web are rigorously studied in cosmology and galaxy physics (de Lapparent et al. 1986; Tempel et al. 2011; Kuutma et al. 2017; Crone Odekon et al. 2018). Galaxy datasets from the Sloan Digital Sky Survey (York et al. 2000), until today the most comprehensive galaxy survey, have been analysed in this thesis. The other physical data analysed in this thesis, the astrophysical turbulent plasma simulation data (Nättilä 2019), describes a phenomena present inside structures of almost all scales in the Universe: galaxy clusters; black hole accretion discs; solar corona and even in the magnetosphere of the Earth. The physical phenomena present in a pool of high-energy particles are not yet fully understood and are studied in various concepts (Cassak & Shay 2007; Retinò et al. 2007; Dupuis et al. 2020). The most comprehensive observational galaxy surveys and state-of-the-art astrophysical plasma simulations give the opportunity to apprehend the unknowns in the field.

Techniques of clustering analysis aim to find sets of objects more similar to each other than they are to objects outside of that set in accordance to a chosen distance metric. The output of the algorithm gives information about the correlations and clusters present in the dataset. Clustering algorithms are developed for a wide set of data regimes and thus provide a comprehensive set of analytical tools. An automated analysis, such as an unsupervised clustering algorithm, detects clusters among input data vectors and learns new unknown physical correlations from the feature space of the data vectors with few or no explicit assumptions. Spatial clustering algorithms are able to detect spatial patterns and associations among sets of events described by their spatial locations.

The thesis applied and also developed a framework for clustering algorithms for the astrophysical datasets. The caveats in the datasets or in applied methodologies were overcome and the clustering algorithms were successfully applied on the catalogues. We analysed two different types of astrophysical data. First of them were datasets about points in an observed region of space, these were the observational galaxy datasets, and catalogues of complex filamentary structure elements detected from the spatial distribution of galaxies (Tempel et al. 2014). The second type of datasets was of astrophysical simulation images, obtained from a particle-in-cell turbulent magnetically dominated collisionless plasma simulation (Nättilä 2019).

These were images consisting of millions of pixels, each of which represented in an $n-$dimensional physical feature space. For spatial objects, we were interested in finding spatial clustering or correlation signals, and for the plasma images, we were interested in finding clusters of similar pixels.

In the first part of this thesis we applied techniques from spatial point pattern theory in order to search for spatial correlation and clustering from the catalogues of cosmic web elements. The techniques of spatial point pattern analysis are well applicable for observational cosmological datasets. The galaxies are just viewed as a realization of a point process (Pons-Bordería et al. 1999; Martinez & Saar 2001; Ripley 1981). The $1-$dimensional spatial correlation analysis technique profusely used in cosmology was able to give information about the galaxy locations inside the cosmic web. The method found regularity in the location of galaxies inside galaxy filaments. The results described the spatial environment of galaxies inside the galaxy filaments, emphasizing the effect of galaxy groups on the surrounding space as matter is drawn towards the group and a gap in galaxy distributions is created. Galaxies and also galaxy groups have a preferred distance in between them, indicating to physical processes driving the merger of galaxies. It also raised questions about the physical mechanisms creating such a pattern. Catalogues of photometric galaxies do not have accurate $3-$dimensional coordinates and the mathematically retrieved network of galaxy filaments possesses most of the structures in a limited distance gap. In the presence of these caveats the bivariate $J-$function was still able to detect spatial clustering between these object catalogues. The method showed clearly whether the new error-prone catalogue of photometric redshift galaxies traces the network of filamentary spines. The result gives basis for the use of these galaxies in the future modelling of the cosmic web. By sampling the filaments dependent on their orientation corresponding to the line line of sight, we estimated the impact of the projection effect and the orientation of filaments on the detected clustering signal.

We showed how the pair correlation function can unravel spatial patterns from catalogues of galaxies and how the bivariate $J-$function is able to detect a positive association or a clustering signal between two projected cosmological object catalogues. These datasets were the photometric redshift galaxies and the galaxy filaments detected with an object point process from the spatial distribution of spectroscopic galaxies. With the growth of datasets and higher accuracy for the objects $3-$dimensional coordinates, more statistically significant results will emerge from the spatial point pattern analysis.

In addition to the spatial point pattern analysis, our aim was to find structures from astrophysical plasma simulation images, whose geometrical properties are of great interest. The Self-Organizing Map was used as the base unsupervised learning algorithm, because the method tries to preserve the topology of the input sam-

ple space and is able to grasp fine structures from an image. The disadvantage of the Self-Organizing Map algorithm is the sensibility to noise and initial conditions. The non-deterministic nature of artificial neural networks was successfully alleviated by combining multiple independent cluster realizations. Combining multiple independent SOM cluster evaluations in an ensemble produced robust estimations for the locations of structures in a magnetically dominated turbulent plasma simulation. Metrics to combine cluster evaluations were created, which enabled to stack independent realizations. These metrics are similar to those widely used in supervised learning, where the obtained result from an algorithm is compared to a previously known ground truth. In our case no such knowledge exists, as the structures are also physically not well understood. The framework produced cluster information for pixels on an image, accurately detecting the most interesting structures in the plasma, the current sheets. The ensemble framework, which stacks multiple unsupervised clustering algorithm evaluations for clusters on an image, produces robust estimates for the boundaries of objects on the image. These boundaries can be successfully used to analyse the geometrical properties of the structures.

Data science has the potential to greatly improve the knowledge pool of other science domains. Among them are clustering techniques, which are applicable for multiple different data regimes. Observational and simulation-generated datasets of astrophysical phenomena have rapidly grown during the past decades and thus they demand powerful and sophisticated analysis techniques. Methods from spatial point patterns and unsupervised machine learning can be fruitfully applied for these datasets, as proven in this thesis.

# References


Baddeley, A., Bárány, I., & Schneider, R. 2007, *Spatial point processes and their applications*, Stochastic Geometry: Lectures given at the CIME Summer School held in Martina Franca, Italy, September 13–18, 2004, 1

Baddeley, A., Rubak, E., & Turner, R. 2015, Spatial point patterns: methodology and applications with R (CRC Press)

Basheer, I. & Hajmeer, M. 2000, *Artificial neural networks: fundamentals, computing, design, and application*, Journal of Microbiological Methods, 43, 3 , neural Computting in Micrbiology

Baxter, E. J. & Rozo, E. 2013, *A Maximum Likelihood Approach to Estimating Correlation Functions*, ApJ, 779, 62

Beck, R., Dobos, L., Budavári, T., Szalay, A. S., & Csabai, I. 2016, *Photometric redshifts for the SDSS Data Release 12*, MNRAS, 460, 1371

Benitez, N., Dupke, R., Moles, M., et al. 2014, *J-PAS: The Javalambre-Physics of the Accelerated Universe Astrophysical Survey*, arXiv e-prints, arXiv:1403.5237

Bond, J. R., Kofman, L., & Pogosyan, D. 1996, *How filaments of galaxies are woven into the cosmic web*, Nature, 380, 603

Bonoli, S., Marín-Franch, A., Varela, J., et al. 2020, *The miniJPAS survey: a preview of the Universe in 56 colours*, arXiv e-prints, arXiv:2007.01910

Brelsford, C., Martin, T., Hand, J., & Bettencourt, L. M. A. 2018, *Toward cities without slums: Topology and the spatial evolution of neighborhoods*, Science Advances, 4

Buryak, O. & Doroshkevich, A. 1996, *Correlation function as a measure of the structure.*, A&A, 306, 1

Cassak, P. A. & Shay, M. A. 2007, *Scaling of asymmetric magnetic reconnection: General theory and collisional simulations*, Physics of Plasmas, 14, 102114

Chang, Y.-M., Baddeley, A., Wallace, J., & Canci, M. 2013, *Spatial statistical analysis of tree deaths using airborne digital imagery*, International Journal of Applied Earth Observation and Geoinformation, 21, 418

Chiu, S. N., Stoyan, D., Kendall, W. S., & Mecke, J. 2013, Stochastic Geometry and its Applications. Third Edition. (John Wiley and Sons)

Colless, M., Dalton, G., Maddox, S., et al. 2001, *The 2dF Galaxy Redshift Survey: spectra and redshifts*, MNRAS, 328, 1039

Comisso, L. & Sironi, L. 2018, *Particle Acceleration in Relativistic Plasma Turbulence*, Phys. Rev. Lett., 121, 255101

Cover, T. & Hart, P. 1967, *Nearest neighbor pattern classification*, IEEE Transactions on Information Theory, 13, 21

Crone Odekon, M., Hallenbeck, G., Haynes, M. P., et al. 2018, *The Effect of Filaments and Tendrils on the H I Content of Galaxies*, ApJ, 852, 142

Cross, R. 2015, Principal Component Analysis Handbook (Clanrye International)

Davis, M., Meiksin, A., Strauss, M. A., da Costa, L. N., & Yahil, A. 1988, *On the Universality of the Two-Point Galaxy Correlation Function*, ApJ, 333, L9

Davis, M. & Peebles, P. J. E. 1983, *A survey of galaxy redshifts. V. The two-point position and velocity correlations.*, ApJ, 267, 465

De Bodt, E., Cottrell, M., & Verleysen, M. 2007, *Statistical tools to assess the reliability of self-organizing maps*, arXiv Mathematics e-prints, math/0701144

de Jong, R. S., Agertz, O., Berbel, A. A., et al. 2019, *4MOST: Project overview and information for the First Call for Proposals*, The Messenger, 175, 3

de Jongh, M. C. & van Lieshout, M. N. M. 2020, *Testing biodiversity using inhomogeneous summary statistics*, arXiv e-prints, arXiv:2007.07635

de Lapparent, V., Geller, M. J., & Huchra, J. P. 1986, *A Slice of the Universe*, ApJ, 302, L1

Ding, S. & Hua, X. 2014, *Recursive least squares projection twin support vector machines for nonlinear classification*, Neurocomputing, 130, 3 , track on Intelligent Computing and Applications Complex Learning in Connectionist Networks

Doguwa, S. I. 1990, *On Edge-Corrected Kernel-Based Pair-Correlation Function Estimators for Point Processes*, Biometrical Journal, 32, 95

Dupuis, R., Goldman, M. V., Newman, D. L., Amaya, J., & Lapenta, G. 2020, *Characterizing Magnetic Reconnection Regions Using Gaussian Mixture Models on Particle Velocity Distributions*, ApJ, 889, 22

Erwin, E., Obermayer, K., & Schulten, K. 1992, *Self-organizing maps: ordering, convergence properties and energy functions*, Biological cybernetics, 67, 47

Fiksel, T. 1988, *Edge-corrected density estimators for point processes*, Statistics, 19, 67

Foxall, R. & Baddeley, A. 2002, *Nonparametric measures of association between a spatial point process and a random set, with geological applications*, Journal of the Royal Statistical Society: Series C (Applied Statistics), 51, 165

Freund, Y. & Schapire, R. E. 1996, *Experiments with a New Boosting Algorithm*, in IN PROCEEDINGS OF THE THIRTEENTH INTERNATIONAL CONFERENCE ON MACHINE LEARNING (Morgan Kaufmann), 148

Gatys, L. A., Ecker, A. S., & Bethge, M. 2015, *A Neural Algorithm of Artistic Style*, arXiv e-prints, arXiv:1508.06576

Hamilton, A. J. S. 1993, *Toward Better Ways to Measure the Galaxy Correlation Function*, ApJ, 417, 19

Hammer, B., Micheli, A., Sperduti, A., & Strickert, M. 2004, *Recursive self-organizing network models*, Neural Networks, 17, 1061

Hansen, L. K. & Salamon, P. 1990, *Neural network ensembles*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 12, 993

Huynh, H. N. 2019, *Spatial point pattern and urban morphology: Perspectives from entropy, complexity, and networks*, Phys. Rev. E, 100, 022320

Illian, J., Penttinen, A., Stoyan, H., & Stoyan, D. 2008, Statistical analysis and modelling of spatial point patterns, Vol. 70 (John Wiley & Sons)

Jõeveer, M., Einasto, J., & Tago, E. 1978, *Spatial distribution of galaxies and of clusters of galaxies in the southern galactic hemisphere*, MNRAS, 185, 357

Jaffe, A., Fetaya, E., Nadler, B., Jiang, T., & Kluger, Y. 2016, *Unsupervised Ensemble Learning with Dependent Classifiers*, in (Cadiz, Spain: PMLR), 351

Kerscher, M., Szapudi, I., & Szalay, A. S. 2000, *A Comparison of Estimators for the Two-Point Correlation Function*, ApJ, 535, L13

Kleinschroth, F., Healey, J. R., Gourlet-Fleury, S., Mortier, F., & Stoica, R. S. 2016, *Effects of logging on roadless space in intact forest landscapes of the Congo Basin*, Conservation Biology, 31, 469–480

Kohonen, T. 1999, *Fast evolutionary learning with batch-type self-organizing maps*, Neural Processing Letters, 9, 153

Kohonen, T. 2001, Self-Organizing maps (Springer-Verlag Berlin Heidelberg)

Kohonen, T. 2013, *Essentials of the self-organizing map*, Neural networks, 37, 52

Kohonen, T. & Mäkisara, K. 1989, *The self-organizing feature maps*, Phys. Scr, 39, 168

Kohonen, T. & Somervuo, P. 2002, *How to make large self-organizing maps for nonvectorial data*, Neural networks, 15, 945

Kuhn, M. A., Baddeley, A., Feigelson, E. D., et al. 2012, *MYStIX First Results: Spatial Structures of Massive Young Stellar Clusters*, arXiv e-prints, arXiv:1208.3492

Kuutma, T., Tamm, A., & Tempel, E. 2017, *From voids to filaments: environmental transformations of galaxies in the SDSS*, A&A, 600, L6

Landrieu, L. & Simonovsky, M. 2018, *Large-Scale Point Cloud Semantic Segmentation with Superpoint Graphs*, in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4558

Landy, S. D. & Szalay, A. S. 1993, *Bias and Variance of Angular Correlation Functions*, ApJ, 412, 64

Lee, J. A. & Verleysen, M. 2002, *Self-organizing maps with recursive neighborhood adaptation*, Neural networks, 15, 993

Lloyd, S. 1982, *Least squares quantization in PCM*, IEEE Transactions on Information Theory, 28, 129

Martinetz, T. & Schulten, K. 1994, *Topology representing networks*, Neural Networks, 7, 507

Martinez, V. J. & Saar, E. 2001, Statistics of the galaxy distribution (CRC Press)

Martinez, V. J. & Saar, E. 2002, *Statistics of galaxy clustering*

Møller, J. & Waagepetersen, R. P. 2004, Statistical inference and simulation for spatial point processes (Chapman and Hall/CRC, Boca Raton)

Mwebaze, E., Bearda, G., Biehl, M., & Zuehlke, D. 2015, *Combining dissimilarity measures for prototype-based classification*, in European Symposium on Artificial Neural Networks (ESANN) 2015, ed. M. Verleysen, Vol. 23 (d-side publishing), 31

Møller, J. & Toftaker, H. 2014, *Geometric Anisotropic Spatial Point Pattern Analysis and Cox Processes*, Scandinavian Journal of Statistics, 41, 414

Møller, J. & Waagepetersen, R. 2007, *Modern Statistics for Spatial Point Processes**, Scandinavian Journal of Statistics, 34, 643

Møller, J. & Waagepetersen, R. 2017, *Some Recent Developments in Statistics for Spatial Point Patterns*, Annual Review of Statistics and Its Application, 4, 317

Nättilä, J. 2019, *Runko: Modern multi-physics toolbox for simulating plasma*, arXiv e-prints, arXiv:1906.06306

Ogata, Y. 1998, *Space-Time Point-Process Models for Earthquake Occurrences*, Annals of the Institute of Statistical Mathematics, 50, 379

Peebles, P. J. E. 1980, The large-scale structure of the universe

Platanios, E. A., Poon, H., Mitchell, T. M., & Horvitz, E. 2017, *Estimating Accuracy from Unlabeled Data: A Probabilistic Logic Approach*, in Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17 (Red Hook, NY, USA: Curran Associates Inc.), 4364–4373

Pons-Bordería, M.-J., Martínez, V. J., Stoyan, D., Stoyan, H., & Saar, E. 1999, *Comparing Estimators of the Galaxy Correlation Function*, ApJ, 523, 480

Povala, J., Virtanen, S., & Girolami, M. 2019, *Burglary in London: Insights from Statistical Heterogeneous Spatial Point Processes*, arXiv e-prints, arXiv:1910.05212

Retinò, A., Sundkvist, D., Vaivads, A., et al. 2007, *In situ evidence of magnetic reconnection in turbulent plasma*, Nature Physics, 3, 236

Ripley, B. D. 1981, Spatial statistics

Rokach, L. 2009, *Collective-agreement-based pruning of ensembles*, Computational Statistics & Data Analysis, 53, 1015

Schneider, P., Biehl, M., & Hammer, B. 2009, *Adaptive Relevance Matrices in Learning Vector Quantization*, Neural Computation, 21, 3532, pMID: 19764875

Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, *The Two Micron All Sky Survey (2MASS)*, AJ, 131, 1163

Snow, J. 1855, On the mode of communication of cholera (John Churchill)

Somerville, R. S., Davis, M., & Primack, J. R. 1997, *A Reanalysis of Small-Scale Velocity Dispersion in the CfA1 Survey*, ApJ, 479, 616

Somervuo, P. & Kohonen, T. 1999, *Self-organizing maps and learning vector quantization for feature sequences*, Neural Processing Letters, 10, 151

Stefanovič, P. & Kurasova, O. 2011, *Influence of learning rates and neighboring functions on self-organizing maps*, in International Workshop on Self-Organizing Maps, Springer, 141

Steinhaus, H. 1956, *Sur la division des corps matériels en parties*, Bull. Acad. Polon. Sci. Cl. III., 4, 801

Stoica, R. S. 2010, *Marked point processes for statistical and morphological analysis of astronomical data*, European Physical Journal Special Topics, 186, 123

Stoica, R. S., Martínez, V. J., & Saar, E. 2010, *Filaments in observed and mock galaxy catalogues*, A&A, 510, A38

Stoyan, D. & Stoyan, H. 1994, Fractals, random shapes, and point fields: methods of geometrical statistics, Vol. 302 (John Wiley & Sons Inc)

Sunyaev, R. A. & Zeldovich, Y. B. 1970, *Small-Scale Fluctuations of Relic Radiation*, Ap&SS, 7, 3

Team, T. M. S., Babusiaux, C., Bergemann, M., et al. 2019, *The Detailed Science Case for the Maunakea Spectroscopic Explorer, 2019 edition*

Tempel, E., Saar, E., Liivamägi, L. J., et al. 2011, *Galaxy morphology, luminosity, and environment in the SDSS DR7*, A&A, 529, A53

Tempel, E., Stoica, R. S., Martínez, V. J., et al. 2014, *Detecting filamentary pattern in the cosmic web: a catalogue of filaments for the SDSS*, MNRAS, 438, 3465

Uritsky, V. M., Pouquet, A., Rosenberg, D., Mininni, P. D., & Donovan, E. F. 2010, *Structures in magnetohydrodynamic turbulence: Detection and scaling*, Phys. Rev. E, 82, 056326

Valada, A., Mohan, R., & Burgard, W. 2019, *Self-Supervised Model Adaptation for Multimodal Semantic Segmentation*, International Journal of Computer Vision, 128, 1239–1285

Valova, I., Georgiev, G., Gueorguieva, N., & Olson, J. 2013, *Initialization issues in self-organizing maps*, Procedia Computer Science, 20, 52

van de Weygaert, R. 1991, *Quasi-periodicity in deep redshift surveys.*, MNRAS, 249, 159

van Lieshout, M. N. M. 2000, Markov Point Processes and their Applications (Imperial College Press, London)

van Lieshout, M. N. M. & Baddeley, A. J. 1996, *A nonparametric measure of spatial interaction in point patterns*, Statistica Neerlandica, 50, 344

van Lieshout, M. N. M. & Baddeley, A. J. 1999, *Indices of dependence between types in multivariate point patterns*, Scandinavian Journal of Statistics, 26, 511

Wang, X., Girshick, R., Gupta, A., & He, K. 2017, *Non-local Neural Networks*

White, S. D. M. 1979, *The hierarchy of correlation functions and its relation to other measures of galaxy clustering.*, MNRAS, 186, 145

York, D. G., Adelman, J., Anderson, John E., J., et al. 2000, *The Sloan Digital Sky Survey: Technical Summary*, AJ, 120, 1579

Zhdankin, V., Werner, G. R., Uzdensky, D. A., & Begelman, M. C. 2017, *Kinetic Turbulence in Relativistic Plasma: From Thermal Bath to Nonthermal Continuum*, Phys. Rev. Lett., 118, 055103

Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., & Torralba, A. 2016, *Learning Deep Features for Discriminative Localization*, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2921

Zhu, Z., Xu, M., Bai, S., Huang, T., & Bai, X. 2019, *Asymmetric Non-Local Neural Networks for Semantic Segmentation*, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 593

# Summary in Estonian

## Astrofüüsikaliste struktuuride uurimine klasteranalüüsi meetoditega

Viimaste aastakümnete jooksul on suuremahulised vaatlused ja arvutisimulatsioonid astrofüüsika arengule plahvatuslikult kaasa aidanud. Hetkel arendamisel olevad suureskaalalised kosmoloogilised vaatlused nagu Maunakea Spectroscopic Explorer (MSE) (Team et al. 2019), Javalambre Physics of the Accelerating Universe Astrophysical Survey (J-PAS) (Benitez et al. 2014) ja VISTA 4-m Multi-Object Spectroscopic Telescope (4MOST) (de Jong et al. 2019) loovad tuleviku teadusavastuste tarbeks veel enam võimalusi. Lisaks vaatlustehnikate arengule on arvutuskeskuste hüppeline areng loonud soodsa pinnase keerukate füüsikaliste protsesside modelleerimiseks numbriliste simulatsioonidega. Antud astronoomilised suureskaalalised vaatlused ja astrofüüsikalised simulatsioonid on toonud endaga kaasa suure kasvu andmete mahus, mis on tõstnud veel olulisemale kohale modernsete optimiseeritud analüüsimeetodite kasutamise ja edasiarendamise.

Antud töö eesmärk oli leida uusi seoseid ja füüsikalisi struktuure astrofüüsikalistest andmestikest. Selleks kasutasime klasteranalüüsi meetodeid ruumilise statistika ja juhendamata masinõppe valdkondadest. Analüüsi tulemuste täpsuse tõstmiseks arendasime välja ka arvutuslikult optimaalse ansambelõppe raamistiku. Eelmainitud meetoditega leidsime korrelatsiooni galaktikate paiknemises Universumi struktuurielementides ja detekteerisime klasterdumise nende struktuurielementide ja uue galaktikate andmestiku vahel. Samuti leidsime juhendamata ansambelõppe abil füüsikalised klastrid astrofüüsikalise plasma andmestikust. Juhendamata klasteranalüüsi abil kaardistasime plasmas leiduvad objektid, mis kirjeldavad plasmas leiduvaid füüsikalisi struktuurielemente. Antud analüüside abil detekteerisime ja aitasime paremini mõista nendes füüsikalistes struktuurides peituvaid fenomene.

Kosmoloogia on teadus, mis uurib meie Universumi tervikuna ning püüab mõista selles sisalduva omavahelisi seoseid ning mõjureid. Tervikliku pildi abil on võimalik kirjeldada Universumi arengut sünnist tema lõpliku saatuseni. Üheks suurimaks astrofüüsikaliseks küsimuseks tänapäeval on lõplikult tõendada ja mõista Universumis paiknevat tumeainet ja tumeenergiat. Universumis leiduv energia on jaotunud järgnevalt – kõigest 4.6% on barüonaine, mis on nähtav elektromagneetilisel spektrumil, 24% on tumeaine, mis on jälgitav ainult tema poolt tekitatud gravitatisooniliste fenomenide abil, ning 71.4% tumeenergia. Tumeaine ja tumeenergia mõistmiseks on tarvilik kaardistada ja uurida nähtavat barüonainet Universumis.

Suurimad nähtavad objektid Universumis on galaktikad, mis moodustavad inimese jaoks hoomamatute mõõtmetega struktuure. See galaktikatest joonistuv struktuur

järgib tumeaine jaotust Universumis. Nähtava aine ja tumeaine jaotust Universumis nimetatakse Universumi suureskaalaliseks struktuuriks. Suureskaalalised galaktikate vaatlused nagu Sloan Digital Sky Survey (SDSS) (York et al. 2000), Two Micron All Sky Survey (2MASS) (Skrutskie et al. 2006) ja 2dF Galaxy Redshift Survey (2dfGRS) (Colless et al. 2001) on loonud tänapäevani kõige suuremad vaatluslikud galaktikate andmestikud. Universumi vaatluslikku struktuuri on intensiivselt uuritud juba aastakümneid (Sunyaev & Zeldovich 1970; Jõeveer et al. 1978; Peebles 1980; Bond et al. 1996). Antud töödest ilmnes Universumi struktuuri keerukus, selle potentsiaalsed füüsikalised põhjused ning mõju selles leiduvatele objektidele. Lisaks eelmainitule on saanud ilmsiks keskkonna mõju galaktikate evolutsioonile (de Lapparent et al. 1986; Tempel et al. 2011; Kuutma et al. 2017; Crone Odekon et al. 2018). Seetõttu omab galaktikate kaardistamine ja antud võrgustiku uurimine kosmoloogias ja galaktikate füüsikas keskset kohta.

Universumis leiduv barüonaine ja tumeaine jaotuvad peamiselt pikkadesse silindrilistesse filamentidesse, mille pikkused ulatuvad mõnest megaparsekist sadade megaparsekiteni, ja sfäärilistesse tihedatesse parvedesse, milles leidub kümneid kuni tuhandeid galaktikaid. Eelnimetatud tihedad keskkonnad piiritlevad peaaegu tühjasid hoomamatu suurusega sfäärilisi tühikuid. Galaktikate vaatlused on tänapäevani teadlasi varustanud kataloogidega, mis sisaldavad enama kui miljoni galaktika spektroskoopilise punanihkeid. Antud andmete abil on võimalik Universumi struktuurielemente kaardistada. Galaktikaliste filamentide andmestik, mida antud töös uuritakse, on detekteeritud just SDSS spektroskoopilise punanihkega galaktikate jaotusest matemaatilise märgistatud punktprotsessi abil (Tempel et al. 2014). Antud töö esimeses pooles uurisime neid pikkasid looklevaid sildasid, mida nimetatakse galaktikalisteks filamentideks. Kosmoloogias on galaktikaid vaadeldud kui punktprotsessi realisatsiooni juba aastakümneid (Pons-Bordería et al. 1999; Martinez & Saar 2001; Ripley 1981). Ka selles doktoritöös kirjeldasime galaktikate paiknemist Universumis kui realisatsiooni punktprotsessist.

Esiteks pakkus meile huvi filamentides leiduvate galaktikate ruumiline jaotus. Selle uurimiseks rakendasime galaktikate paiknemise andmetele paariskorrelatsioonifunktsiooni. Antud füüsikaline korrelatsioonimõõt annab meile kirjelduse galaktikatest moodustunud punktmustris leiduvale ruumilisele korrelatsioonile. Me selgitasime välja, et galaktikalistes filamentides leidub kindel muster galaktikate paiknemises. Nimelt paiknevad galaktikad eelistatult teineteisest umbes $7.5 - 8.0\,\mathrm{h}^{-1}$Mpc kaugusel ning galaktikaparvede ümbrused on umbes $2\,\mathrm{h}^{-1}$Mpc raadiuses tühjad. Antud tulemus viitab füüsikalistele protsessidele, mis kontrollivad galakatikate evolutsiooni antud struktuurielemendis.

Teiseks pakkus meile huvi fotomeetrilise vaatluse käigus omandatud galaktikate andmestik. Antud fotomeetrilised punanihked on väga halva määramistäpsusega,

mille tõttu pole nende galaktikate 3–mõõtmelised koordinaadid usaldusväärsed ning seal leiduvat infot pole antud filamentstruktuuri modelleerimisel kasutatud. Nendes kataloogides leidub aga infot miljonite galaktikate ruumilise asukoha kohta ning nad võivad oluliselt panustada filamentstruktuuri modelleerimisse. Fotomeetrilise punanihkega andmestike potensiaalse panuse välja selgitamiseks kasutasime ruumilise statistika klasteranalüüsi meetodit – kahemuutuja $J-$funktsiooni. Antud ruumilise klasterdumise hindamise statistik näitas selget klasterdumise signaali spektroskoopilistest galaktikatest modelleeritud filamentide ja fotomeetriliste galaktikate vahel. See tõendas meile, et fotomeetriliste galaktikate andmestikus leidub palju olulist infot filamentstruktuuri kohta. Samuti tõendasid need galaktikad modelleeritud filamentide õigsust ning andmete filtreerimise abil hindasime erinevate projektsiooniefektide mõju.

Lisaks kosmoloogilistele andmestikele analüüsisime antud töös astrofüüsikalise plasma simulatsiooni andmeid. Nimelt uurisime turbulentse põrkevaba magneetiliselt domineeritud plasma simulatsioon (Nättilä 2019). Antud simulatsioon kirjeldab füüsikalist fenomeni, mis eksisteerib väga erineva suurusega Universumi elementides. Hõre magneetiliselt domineeritud laetud osakeste plasma eksisteerib galaktikate parvedes, mustade aukude akretsiooniketastes, kahe neutrontähe kokkupõrke keskkonnas, Päikese koroonas, Maa magnetosfääris ning isegi maapealsetes tuumasünteesi reaktorites. Antud plasma uurimine omab seega olulist rolli nii kosmoloogiliste objektide mõistmisel kui ka inimeste elukeskkonna mõistmises ja arendamises. Tänapäevani ei ole astrofüüsikalises plasmas leiduvad füüsikalised fenomenid täielikult välja selgitatud ning selles leiduvate struktuuride automaatne detekeerimine on plasmafüüsikas veel täielikku lahendamist vajav probleem (Zhdankin et al. 2017; Comisso & Sironi 2018; Nättilä 2019; Cassak & Shay 2007; Retinò et al. 2007; Dupuis et al. 2020).

Antud töös rakendasime turbulentse põrkevaba magneetiliselt domineeritud plasma simulatsiooni (Nättilä 2019) andmetest saadud piltidele juhendamata klasteranalüüsi meetodit nimega *Self-Organizing Map*. Eelmainitud meetod kaardistab plasmas leiduvad füüsikalised struktuurid. Antud tulemus ei ole aga geomeetrilise ja füüsikalise analüüsi tarbeks piisavalt usaldusväärne, sest juhendamata närvivõrk on stohhastiline ning ei saavutanud globaalset optimumi. Seega arendasime ja rakendasime plasma andmetele juhendamata ansambelõpet, mis leidis täpsemad struktuuride klassifikatsioonid. Kasutades juhendamata ansambelõpet klassifitseerisime piksel-haaval plasma simulatsiooni pildid ning seeläbi detekteerisime plasmas leiduvad füüsikalised struktuurid. Saadud struktuuride edasine füüsikaline ja geomeetriline analüüs on suure tähtsusega ning annab selgust antud struktuurides leiduvatele fenomenidele.

Antud doktoritöös rakendasime me erinevaid klasteranalüüsi meetodeid astrofüüsikaliste struktuuride uurimiseks ja detekteerimiseks. Meetodid olid ruumilise sta-

tistika ja juhendamata masinõppe valdkondadest. Leitud muster galaktikate paiknemises piki filamente kirjeldab filamentide keskkonda ning nende mõju galaktika evolutsioonile. Fotomeetriliste galaktikate selge klasterdumine leitud filamentide telgedega motiveerib tulevikus antud galaktikaid modelleerimisel kasutama. Fotomeetriliste galaktikate kasutamine filamentstruktuuride modelleerimisel võimaldab kaardistada detailsema ja kaugemale ulatuvama universumi struktuuri. Hetkel arendamisel olev vaatlustprojekt Javalambre Physics of the Accelerating Universe Astrophysical Survey (J-PAS) (Benitez et al. 2014) kaardistab miljonite galaktikate fotomeetrilise punanihke oluliselt suurema täpsusega. J-PAS galaktikate fotomeetrilisi punanihkeid on plaanis tulevases teadustöös Universumi struktuuri modellerimises kasutada.

Füüsikaliste struktuuride kaardistamine astrofüüsikalise plasma simulatsioonis andis automatiseeritud meetodi siiani põhjalikult mõistmata elementide leidmiseks. Saadud andmestik avab tee turbulentses plasmas toimuvate protsesside mõistmiseks. Kõige enam huvipakkuvad elemendid neist on *current sheets*, mille läbi kõrgelt laetud osakesed plasma keskkonnast välja pääsevad. Eelmainitud struktuuride geomeetriline analüüs plasma simulatsiooni piltidest on hetkel käsilolev töö. Turbulentne põrkevaba magneetiliselt domineeritud plasma on simuleeritud kasutades superarvuteid. Järgmise aasta keskel hakkab 10 Euroopa riigi konsortsiumi juhtimisel tööle võimsuselt maailma tippude hulka kuuluv superarvuti LUMI[1]. Seega täpsustuvad ja suurenevad ka mitmed astrofüüsikaliste fenomenide simulatsioonid, sealhulgas eelmainitud plasma simulatsioon. See toob endaga kaasa andmete hulga kasvu ning muudab optimeeritud ja paralleliseeritud analüüsimeetodite kasutamise veelgi olulisemaks. Antud töös kasutatud ja arendatud juhendamata ansambelõppe raamistik on lihtsasti paralleliseeritav ja kasutab analüüsimisel arvutuslikult kiireid operatsioone. See tõstab antud metoodikate kasutamise olulisust ka tuleviku simulatsooniandmete uurimisel.

Järgmiste aastate jooksul muutuvad paljud antud töös käsitletud andmestikud suuremahulisemaks ja täpsemaks. Antud töös arendatud metoodikad on suureks abiks, et nendest andmetest leida uusi seoseid ja vastata juba olemasolevatele teaduslikele küsimustele.

[1] www.lumi-supercomputer.eu

## ACKNOWLEDGEMENTS

During my PhD period I have met many people who have motivated me to do science and helped me with problems I was struggling with. It would be difficult to name all of these people separately.

Specifically, I want to thank the whole cosmology and galaxy physics team in the Observatory, all of you made me feel welcome and appreciated from the beginning of my PhD studies, as I was not a physicist by training but a statistician. You always made me feel welcome and helped me understand the mysteries of physics.

I am deeply grateful to my supervisors Elmo Tempel and Radu Stoica, who were my mentors in cosmology and statistics. Especially, I am very thankful to Elmo for teaching me about the science of cosmology and always being considerate and helpful. I appreciate greatly and am very thankful for Radu's scientific impact on my work and his advice about the more political side of science. Elmo and Radu provided me with the possibility to work independently with my collaborator, which I appreciate enormously. I want to give my appreciation to my collaborator Joonas Nättilä, who introduced me to the world of plasma physics and mentored me about science and physics.

I want to give big thanks to Rain Kipper whose office door was always open for all of my questions related to science or not, and providing humorous comments if the situation acquired it. I want to thank Peeter Tenjes for helping me study for the doctoral exams, his door was always open for questions and solving problems on the white board. I want to thank Indrek Vurm for helping me understand magnetic field theory and all the discussions about plasma physics. I wish to thank Antti Tamm for scientific discussions and for bringing a sense of joy to the team and Juhan Liivamägi for helping me with all science or technology related problems. I am also thankful to Jukka Nevalainen for showing an example for how to be passionate about science. I would like to thank Maret Einasto, Enn Saar and Jaan Einasto for all the nice discussions and motivating me to do science.

A crucial driving force for the whole PhD era is of course the whole team of PhD students. A very important person from the clique was my office mate and scientific sister Punyakoti Ganeshaiah Veena. We created a beautiful team and we always supported each other. The PhD team wouldn't have been the same without Toni Tuominen, who always made us laugh and listened to all of our ideas. I will miss our drives to the Observatory every morning, where we dissected all the issues in the political and social scene. I also want to thank Moorits Mihkel Muru, Teet Kuutma and Daniel Blixt for all the good and interesting times and scientific discussions in Tartu and the Observatory. I also want to thank Heleri Ramler for all our discussions

and talks in your office.

I am thankful for my supervisor Radu Stoica and Elmo Tempel for providing me with fruitful environments during my visits to the Université de Lorraine in France and Leibniz Institute for Astrophysics Potsdam (AIP) in Germany. I am grateful for Joonas Nättilä for making me feel warmly welcome in my research visits to Nordic Institute for Theoretical Physics in Sweden.

I gratefully acknowledge the financial support by the Estonian Science Foundation and the Estonian Ministry of Education.

I would like to thank my parents Merle and Igor Bussov, who have always supported me in my endeavours and encouraged me and my siblings to be curious and imaginative. I am thankful to my sister Kadri Bussov and brother Karl Bussov for keeping me on my toes and always supporting me. I also want to thank my grandmother Lilian Volmer for always asking questions about my work and encouraging me along the way.

Last but definitely not least I want to thank my partner Pekka Manninen, who mentored and pushed me through the last year of my PhD. Your comments on my thesis and on my scientific work have been of immense value. I am very grateful for the belief you have in me.

# PUBLICATIONS

## Curriculum vitae

### Personal data


| | |
|---|---|
| Name | Maarja Bussov |
| Date and place of birth | 23 March 1989, Tartu, Estonia |
| Citizenship | Estonian |
| Current employment | Helsinki University (Doctoral Student) |
| Address | Physicum<br>FI-00014 University of Helsinki<br>Finland |
| Phone | (+372) 5834 0926 |
| E-mail | mbussov@ut.ee |


### Education

| | |
|---|---|
| 2005 – 2008 | Nõo Science Gymnasium |
| 2008 – 2011 | University of Tartu, undergraduate student,<br>BSc 2011 (mathematical statistics) |
| 2011 – 2014 | University of Tartu, graduate student,<br>MSc 2014 (mathematical statistics) |
| 2016 – 20xx | University of Tartu, PhD student,<br>Ph.D. xx (physics) |

### Employment

| | |
|---|---|
| 2014 – 2014 | Tartu Observatory, Engineer |
| 2015 – 2016 | TransferWise Ltd, Anti-fraud analyst |
| 2016 – 2020 | University of Tartu, Tartu observatory,<br>Junior researcher |
| 2020– . . . | University of Helsinki, Department of Physics,<br>Doctoral student |

### Professional training

| | |
|---|---|
| 24.10 – 28.10 2016 | Workshop “Dark Matter in the Era of Gaia”,<br>Nordita, Sweden |
| 15.05 – 19.05.2017 | Summer School “Galaxy Formation and Evolution<br>in a Cosmological Context", Spetses, Greece |

| | |
|---|---|
| 27.06 – 05.07.2017 | Summer School “CSC Summer School in HPC", Nuuksio, Finland |
| 09.10 – 13.10.2017 | Summer School “School of Statistics for Astrophysics 2017: Bayesian Methodology", Autrans, France |
| 26.03 – 28.03.2018 | Course “Advanced Fortran Programming", CSC, Finland |
| 20.05 – 25.05.2018 | Summer School “AIDA IX - Statistical Challenges in 21st Century Cosmology", Valencia, Spain |
| 04.11 – 10.11.2018 | Winter School “Canary Islands Winter School of Astrophysics: Big Data Analysis in Astronomy", Tenerife, Spain |

**Conference presentations**

| | |
|---|---|
| 1.10 – 3.10 2014 | Conference "‘Tuorla – Tartu annual meeting 2014: Small and Large scale Universe"’, Tuorla, Finland<br>*Oral presentation*: "Galaxy filaments as pearl necklaces" |
| 26.09 – 29.09 2017 | Conference "‘Tuorla – Tartu annual meeting 2017: What matters?", Tuorla, Finland<br>*Oral presentation*: "The contribution of SDSS photometric redshift galaxies to the defined filamentary structure elements of the Universe" |
| 03.10 – 05.10 2018 | Workshop “Tartu – Tuorla annual meeting 2018: The large scale properties of the universe as a whole", Trofee hunting lodge, Estonia<br>*Oral presentation*: “Photometric redshift galaxies as tracers of the filamentary network” |
| 22.05 – 25.05.2018 | Conference “Cosmo 21: Statistical challenges in the 21st century cosmology", Valencia, Spain<br>*Poster*: “Photometric redshift galaxies as tracers of the filamentary network” |
| 05.03 – 07.03 2019 | Conference “Finnish Physics Days", Helsinki, Finland<br>*Oral presentation*: “Photometric redshift galaxies as tracers of the filamentary network” |
| 03.06 – 07.06 2019 | Conference “51es Journées de Statistique 2019", Nancy, France<br>*Oral presentation*: “The bivariate J-function to analyse positive association between galaxies and galaxy filaments” |

## Language skills

| | |
|---|---|
| Estonian | the first language |
| English | very good |
| German | lower mid-level |
| Finnish | lower mid-level |

## Honours and Awards

| | |
|---|---|
| 2019 | Nordita Visiting Ph.D. Fellow<br>(Nordic Institute of Theoretical Physics, Sweden) |
| 2019 | HPC-Europa3 Transnational Access programme grant<br>(the European Commission Horizon 2020) |
| 2019 | DoRa Stipend for research visit in Institut Elie Cartan<br>de Lorraine, Nancy, France<br>(Foundation Archimedes, Estonia) |
| 2018 | Kristjan Jaak scholarship for short study visits abroad<br>(Estonian Ministry of Education and Research;<br>Foundation Archimedes, Estonia) |
| 2017 | Julius Ernst Öpik scholarship in Astronomy<br>(Tartu Observatory, Estonia) |
| 2017 | DoRa stipend for school in Institut de Planétologie<br>et d'Astrophysique de Grenoble<br>(Foundation Archimedes, Estonia) |

## Fields of research

Large-scale structure of the Universe, spatial point patterns,
turbulent plasma, machine learning.

## Publications

1. **Bussov, M.** and Nättilä, J. 2020, *Segmentation of turbulent computational fluid dynamics simulations with unsupervised ensemble learning*, in preparation
2. **Kruuse, M.**, Tempel, E., Kipper, R., and Stoica, R. S. 2019, *The bivariate J-function to analyse positive association between galaxies and galaxy filaments*, 51es Journées de Statistique 2019, Nancy, France
3. **Kruuse, M.**, Tempel, E., Kipper, R., and Stoica, R. S. 2019, *Photometric redshift galaxies as tracers of the filamentary network*, Astronomy & Astrophysics Journal, 625, A130

4. Tempel, E., **Kruuse, M.**, Kipper, R., Tuvikene, T., Sorce, J. G., and Stoica, R. S. 2018, *Bayesian group finder based on marked point processes. Method and feasibility study using the 2MRS data set*, Astronomy & Astrophysics Journal, 618, A8.
5. Tempel, E. and **Bussov, M.** 2016, *Filamentary pattern in the cosmic web: galaxy filaments as pearl necklaces*, Proceedings of the International Astronomical Union (236 – 241), Cambridge University Press.
6. Tempel, E., Kipper, R., Saar, E., **Bussov, M.**, Hektor, A. and Pelt, J. 2014, *Galaxy filaments as pearl necklaces*, Astronomy & Astrophysics Journal, 572, A8.

# ELULOOKIRJELDUS

## Isikuandmed


| | |
|---|---|
| Nimi | Maarja Bussov |
| Sünniaeg ja -koht | 23. märts 1989, Tartu, Eesti |
| Kodakondsus | eesti |
| Praegune töökoht | Helsinki Ülikool (doktorant) |
| Aadress | Füüsika osakond |
| Address | Physicum |
| | FI-00014 University of Helsinki |
| | Finland |
| E-mail | mbussov@ut.ee |


## Haridus

| | |
|---|---|
| 2005 – 2008 | Nõo Reaalgümnaasium |
| 2008 – 2011 | Tartu Ülikool, üliõpilane, BSc 2003 (matemaatiline statistika) |
| 2011 – 2014 | Tartu Ülikool, magistrant, MSc 2005 (matemaatiline statistika) |
| 2016 – 20xx | Tartu Ülikool, doktorant (füüsika) |

## Teenistuskäik

| | |
|---|---|
| 2014 – 2014 | Tartu Observatoorium, insener |
| 2015 – 2016 | TransferWise Ltd, petturlusevastane analüütik |
| 2016 – 2020 | Tartu Ülikool, Tartu observatoorium, nooremteadur |
| 2020 – . . . | Helsinki Ülikool, Füüsika osakond, doktorant |

## Täiendkoolitus

| | |
|---|---|
| 24.10 – 28.10 2016 | Töökoda “Dark Matter in the Era of Gaia”, Nordita, Rootsi |
| 15.05 – 19.05.2017 | Suvekool “Galaxy Formation and Evolution in a Cosmological Context", Spetses, Kreeka |
| 27.06 – 05.07.2017 | Suvekool “CSC Summer School in HPC", Nuuksio, Soome |
| 09.10 – 13.10.2017 | Suvekool “School of Statistics for Astrophysics 2017: Bayesian Methodology", Autrans, Prantsusmaa |

| | |
|---|---|
| 26.03 – 28.03.2018 | Kursus “Advanced Fortran Programming", CSC, Soome |
| 20.05 – 25.05.2018 | Suvekool “AIDA IX - Statistical Challenges in 21st Century Cosmology", Valencia, Hispaania |
| 04.11 – 10.11.2018 | Talvekool “Canary Islands Winter School of Astrophysics: Big Data Analysis in Astronomy", Tenerife, Hispaania |

**Konverentside ettekanded**

| | |
|---|---|
| 1.10 – 3.10 2014 | Konverents „Tuorla – Tartu annual meeting 2014: Small and Large scale Universe“, Tuorla, Soome<br>*Suuline ettekanne*: „Galaxy filaments as pearl necklaces" |
| 26.09 – 29.09 2017 | Konverents „Tuorla – Tartu annual meeting 2017: What matters?“, Tuorla, Soome<br>*Suuline ettekanne*: "The contribution of SDSS photometric redshift galaxies to the defined filamentary structure elements of the Universe" |
| 03.10 – 05.10 2018 | Konverents “Tartu – Tuorla annual meeting 2018: What matters?", Trofee Jahimaja, Eesti<br>*Suuline ettekanne*: “Photometric redshift galaxies as tracers of the filamentary network” |
| 22.05 – 25.05.2018 | Konverents “Cosmo 21: Statistical challenges in the 21st century cosmology", Valencia, Hispaania<br>*Plakat*: “Photometric redshift galaxies as tracers of the filamentary network” |
| 05.03 – 07.03 2019 | Konverents “Finnish Physics Days", Helsinki, Soome<br>*Suuline ettekanne*: “Photometric redshift galaxies as tracers of the filamentary network” |
| 03.06 – 07.06 2019 | Konverents “51es Journées de Statistique 2019", Nancy, Prantsusmaa<br>*Suuline ettekanne*: “The bivariate J-function to analyse positive association between galaxies and galaxy filaments” |

**Keelteoskus**

| | |
|---|---|
| eesti keel | emakeel |
| inglise keel | väga hea |
| saksa keel | madal kesktase |
| soome kee | madal kesktase |

**Uurimistoetused ja stipendiumid**

| | |
|---|---|
| 2019 | Nordita Visiting Ph.D. Fellow<br>(Nordic Institute of Theoretical Physics, Rootsi) |
| 2019 | HPC-Europa3 Transnational Access programme grant<br>(European Commission Horizon 2020) |
| 2019 | DoRa stipendium lühiajaliseks visiidiks,<br>Institut Elie Cartan de Lorraine, Nancy, France |
| 2018 | Kristjan Jaak stipendium lühikesteks visiitideks<br>(Tartu Ülikool) |
| 2017 | Julius Ernst Öpik stipendium<br>(Tartu Observatoorium) |
| 2017 | DoRa stipendium teaduskoolis osalemiseks, Institut<br>de Planétologie et d'Astrophysique de Grenoble<br>(Foundation Archimedes, Estonia) |

**Peamised uurimissuunad**

Universumi suureskaalaline struktuur, punkt protsessid,
turbulentne astrofüüsikaline plasma, masinõpe

# DISSERTATIONES ASTRONOMIAE UNIVERSITATIS TARTUENSIS

1. **Tõnu Viik.** Numerical realizations of analytical methods in theory of radiative transfer. Tartu, 1991.
2. **Enn Saar.** Geometry of the large scale structure of the Universe. Tartu, 1991.
3. **Maret Einasto.** Morphological and luminosity segregation of galaxies. Tartu, 1991.
4. **Urmas Haud.** Dark Matter in galaxies. Tartu, 1991.
5. **Eugene A. Ustinov.** Inverse problems of radiative transfer in sounding of planetary atmospheres. Tartu, 1992.
6. **Peeter Tenjes.** Models of regular galaxies. Tartu, 1993.
7. **Ivar Suisalu.** Simulation of the evolution of large scale structure elements with adaptive multigrid method. Tartu, 1995.
8. **Teimuraz Shvelidze.** Automated quantitative spectral classification of stars by means of objective prism spectra: the method and applications. Tartu, 1999.
9. **Jelena Gerškevitš.** Formation and evolution of binary systems with compact objects. Tartu, 2002.
10. **Ivan Suhhonenko**. Large-scale motions in the universe. Tartu, 2003.
11. **Antti Tamm.** Structure of distant disk galaxies. Tartu, 2006.
12. **Vladislav-Veniamin Pustynski.** Modeling the reflection effect in pre-cataclysmic binary systems. Tartu, 2007.
13. **Anna Aret.** Evolutionary separation of mercury isotopes in atmospheres of chemically peculiar stars. Tartu, 2009.
14. **Mari Burmeister.** Characteristics of the hot components of symbiotic stars. Tartu, 2010.
15. **Elmo Tempel.** Tracing galaxy evolution by their present-day luminosity function. Tartu, 2011.
16. **Anti Hirv.** Estimation of time delays from light curves of gravitationally lensed quasars. Tartu, 2011.
17. **Rain Kipper**. Galaxy modelling: dynamical methods and applications. Tartu, 2016, 134 p.
18. **Lauri Juhan Liivamägi**. Properties and spatial distribution of galaxy superclusters. Tartu, 2017, 185 p.
19. **Jaan Laur**. Variability survey of massive stars in Milky Way star clusters. Tartu, 2017, 183 p.
20. **Boris Zhivkov Deshev.** On the coevolution of galaxies and their host clusters. Tartu, 2019, 199 p.
21. **Tiina Liimets**. Nebulosities and jets from outbursting evolved stars. Tartu, 2019, 207 p.