

# **A Data-Driven Approach for Generating Vortex Shedding Regime Maps for an Oscillating Cylinder**

by

Matthew Cann

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Applied Science

in

Mechanical and Mechatronics Engineering

Waterloo, Ontario, Canada, 2022

© Matthew Cann 2022

### **Author's Declaration**

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

## Abstract

Recent developments in wind energy extraction methods from vortex-induced vibration (VIV) have fueled the research into vortex shedding behaviour. The vortex shedding map is vital for the consistent use of normalized amplitude and wavelength to validate the predicting power of forced vibration experiments. However, there is a lack of demonstrated methods of generating this map at Reynolds numbers feasible for energy generation due to the high computational cost and complex dynamics.

Leveraging data-driven methods addresses the limitations of the traditional experimental vortex shedding map generation, which requires large amounts of data and intensive supervision that is unsuitable for many applications and Reynolds numbers. This thesis presents a data-driven approach for generating vortex shedding maps of a cylinder undergoing forced vibration that requires less data and supervision while accurately extracting the underlying vortex structure patterns.

The quantitative analysis in this dissertation requires the univariate time series signatures of local fluid flow measurements in the wake of an oscillating cylinder experiencing forced vibration. The datasets were extracted from a 2-dimensional computational fluid dynamic (CFD) simulation of a cylinder oscillating at various normalized amplitude and wavelength parameters conducted at two discrete Reynolds numbers of 4000 and 10,000. First, the validity of clustering local flow measurements was demonstrated by proposing a vortex shedding mode classification strategy using supervised machine learning models of random forest and  $k$ -nearest neighbour models, which achieved 99.3% and 99.8% classification accuracy using the velocity sensors orientated transverse to the pre-dominant flow ( $u_y$ ), respectively. Next, the dataset of local flow measurement of the  $y$ -component of velocity was used to develop the procedure of generating vortex shedding maps using unsupervised clustering techniques. The clustering task was conducted on subsequences of repeated patterns from the whole time series extracted using the novel matrix profile method. The vortex shedding map was validated by reproducing a benchmark map produced at a low Reynolds number. The method was extended to a higher Reynolds number case of vortex shedding and demonstrated the insight gained into the underlying dynamical regimes of the physical system. The proposed multi-step clustering methods denoted Hybrid Method B, combining Density-Based Clustering Based on Connected Regions with High Density (DBSCAN) and Agglomerative algorithms, and Hybrid Method C, combining  $k$ -Means and Agglomerative algorithms demonstrated the ability to extract meaningful clusters from more complex vortex structures that become increasingly indistinguishable. The data-driven methods yield exceptional performance and versatility, which significantly improves the map generation method while reducing the data input and supervision required.

## **Acknowledgements**

First and foremost, I would like to thank my supervisors, Professor Fue-Sang Lien and Professor William Melek, for their guidance throughout my master's career as I began my journey into the data science field. I am grateful to the ever-knowledgeable Eugene Yee for providing valuable feedback and comments on my work over the years.

I owe many thanks to Ryley McConkey for providing the computational fluid dynamic simulation data for which this research would not be possible.

I would also like to thank my friends and family, who have supported me throughout my academic career. A special thanks to my roommates, Daniel Sola, Andrew Downie, and Aidan Keaveny, for countless stimulating lunch walks and entertainment throughout the two years. Additional thanks to Davis McClarty for the motivation and laughs during all of university.

Lastly, I acknowledge the support of the University of Waterloo, which allowed my academic journey to be possible.

# Table of Contents

List of Figures .....	viii
List of Tables .....	xi
Chapter 1 Introduction.....	1
1.1 Objectives .....	2
1.2 Contributions .....	3
1.3 Thesis Organization .....	3
Chapter 2 Background.....	4
2.1 Introduction.....	4
2.2 Vortex-Induced Vibration.....	4
2.2.1 Theory.....	4
2.2.2 Free and Forced Vibration .....	8
2.2.3 Vortex Shedding Modes .....	9
2.2.4 Vortex Shedding Modes at High Reynolds Number .....	11
2.3 Time Series Clustering .....	13
2.3.1 Subsequence Time Series Clustering.....	14
2.3.2 Cluster Analysis.....	15
2.3.3 Time Series Representation .....	15
2.3.4 Similarity Measure.....	16
2.3.5 Clustering Algorithm .....	19
2.3.6 Evaluation Metric .....	22
Chapter 3 Literature Review .....	25
3.1 Vortex Shedding Map Generation .....	25
3.2 Data-Driven Methods for Vortex Shedding.....	26
3.2.1 Vortex Shedding Mode Classification.....	26
3.2.2 Vortex Shedding Mode Clustering .....	27
3.3 Summary.....	28

Chapter 4	Methodology.....	30
4.1	Datasets.....	30
4.1.1	Dataset A .....	32
4.1.2	Dataset B.....	33
4.2	Mode Classification .....	33
4.2.1	Preprocessing.....	33
4.2.2	Machine Learning Models .....	35
4.3	Clustering Analysis.....	35
4.3.1	Preprocessing.....	35
4.3.2	Streamwise Analysis.....	36
4.3.3	Subsequence Data Mining .....	38
4.3.4	Proposed Clustering Methods.....	39
Chapter 5	Mode Classification using Machine Learning .....	43
5.1	Machine Learning Model Performance .....	43
5.2	Feature Noise Analysis .....	45
5.3	Summary.....	47
Chapter 6	Vortex Shedding Map Generation at Low Reynolds Number.....	48
6.1	Methodology.....	48
6.1.1	Data Exploration.....	49
6.1.2	Subsequence Extraction.....	50
6.1.3	Number of Clusters.....	51
6.2	Proposed Traditional Clustering Methods .....	51
6.2.1	<i>k</i> -Means .....	51
6.2.2	Agglomerative .....	54
6.2.3	Discrete Cosine Transform Representation with <i>k</i> -Means.....	57
6.3	Proposed Hybrid Clustering Methods.....	60
6.3.1	Hybrid Method A.....	60
6.3.2	Hybrid Method B.....	63
6.3.3	Hybrid Method C.....	66
6.4	Discussion.....	70
6.4.1	Traditional Methods.....	70
6.4.2	Hybrid Methods .....	73
6.5	Summary.....	76
Chapter 7	Vortex Shedding Map at High Reynolds Number.....	78
7.1	Methodology.....	78

7.1.1	Data Exploration.....	78
7.1.2	Subsequence Extraction.....	79
7.2	Proposed Traditional Clustering Methods.....	80
7.2.1	<i>k</i> -Means.....	80
7.2.2	Agglomerative.....	83
7.2.3	DCT Time Series Representation with <i>k</i> -Means.....	86
7.3	Proposed Hybrid Clustering Methods.....	88
7.3.1	Hybrid Method A.....	88
7.3.2	Hybrid Method B.....	91
7.3.3	Hybrid Method C.....	94
7.4	Discussion.....	97
7.5	Summary.....	100
Chapter 8	Conclusions and Future Work.....	101
8.1	Future work.....	102
References	.....	103

# List of Figures

Figure 1.1. Helical strakes on smoke towers for the mitigation of VIV [6].	1
Figure 1.2. Forced vibration vortex shedding map in normalized amplitude–wavelength space [7].	2
Figure 2.1. Illustration of a cylinder in free stream and sinusoidal vertical motion.	4
Figure 2.2. Strouhal Number and Reynolds Number relationship [9].	6
Figure 2.3. Modelled cylinder for transverse oscillations in crossflow.	8
Figure 2.4. Free vibration and forced vibration in fluid-structure energy transfer regions. [10].	9
Figure 2.5. Vortex shedding map in the normalized amplitude–wavelength plane for forced vibration [7].	10
Figure 2.6. Point vortex models for the wake regimes of (a) 2S, (b) 2P, and (c) 2PO, (d) P+S, (e) 2P+2S.	13
Figure 2.7. Taxonomy of time series clustering approaches.	13
Figure 2.8. Example of a motif discovered in original time series.	14
Figure 2.9. Taxonomy of cluster analysis.	15
Figure 2.10. Signal representation and decomposition in the time and frequency domain.	16
Figure 2.11. Taxonomy of distance measures.	16
Figure 2.12. Differences in distance measurements between Euclidean and Dynamic Time Warping [36].	19
Figure 2.13. Illustration of the clustering representation of (a) $k$ -Means and (b) $k$ -Medoids.	20
Figure 2.14. Clustering phases implemented in the two-phase method [44].	21
Figure 2.15. Illustration of the inertia metric for a cluster.	22
Figure 2.16. Illustration of the silhouette coefficient for two clusters.	23
Figure 2.17. Illustration of the Dunn index for two clusters.	24
Figure 4.1. Domain and boundary conditions used for the 2D simulation (Not to scale).	31
Figure 4.2. CFD sampling lines in the cylinder wake.	31
Figure 4.3. Sample vorticity colour maps and $x$ -velocity component sensor signals for the three wake modes: 2S (top), 2P, (middle), and 2PO (bottom).	32
Figure 4.4. Frequency spectrum for (a) 2S, (b) 2P, and (c) 2PO vortex shedding modes.	34
Figure 4.5. Frequency domain feature vector given by the Gaussian fit parameters.	35
Figure 4.6. Sensor variance of streamwise sampling lines for 2S modes at $(\lambda^*, A^*) = (4, 0.1)$ for $Re = 4000, 10,000$ .	36
Figure 4.7. Sensor variance of streamwise sampling lines for 2P modes at $(\lambda^*, A^*) = (6, 0.3)$ for $Re = 4000$ .	37
Figure 4.8. Sensor variance of streamwise sampling lines for 2P modes $(\lambda^*, A^*) = (6, 0.3)$ for $Re = 10,000$ .	37
Figure 4.9. Matrix profile example.	38
Figure 4.10. Distance matrix to obtain matrix profile, modified from [72].	39



Figure 4.11. Block diagram for the proposed hybrid algorithms.....	40
Figure 4.12. Example of the hybrid clustering procedure. ....	41
Figure 5.1. Frequency feature space of vortex shedding classes using (a) flow Speed, $u$ , (b) $x$ velocity component, $u_x$ , (c) $y$ velocity component, $u_y$ , and (d) vorticity, $\omega$ . ....	43
Figure 5.2. Cross-validation scores for each of the local measurement sensor data used to train a variety of machine learning models. ....	44
Figure 5.3. Feature noise testing accuracy results for classifiers (a) Decision tree, (b) Random Forest, (c) MLP, and (d) $k$ -NN ....	46
Figure 6.1. Dataset sampled nodes overlaid on reference normalized amplitude–wavelength plane [7].	49
Figure 6.2. Clusters associated with vortex shedding map labels.....	50
Figure 6.3. Example of motif extraction for signals that resemble (a) 2S and (b) 2PO.....	51
Figure 6.4. Generated clusters by $k$ -Means method at $Re = 4000$ .....	52
Figure 6.5. Cluster $t$ -SNE distribution at $Re = 4000$ using $k$ -Means. ....	53
Figure 6.6. Vortex shedding map using $k$ -Means method at $Re = 4000$ . ....	54
Figure 6.7. Generated clusters by Agglomerative (complete, cosine) method at $Re = 4000$ . ....	55
Figure 6.8. Agglomerative dendrogram at level $p = 3$ . ....	55
Figure 6.9. Cluster $t$ -SNE distribution using the agglomerative method at $Re = 4000$ . ....	56
Figure 6.10. Vortex shedding map using the agglomerative method at $Re = 4000$ . ....	57
Figure 6.11. Generated clusters by $k$ -Means method on DCT dataset at $Re = 4000$ .....	58
Figure 6.12. Cluster $t$ -SNE distribution at $Re = 4000$ using $k$ -Means on DCT dataset. ....	58
Figure 6.13. Vortex shedding map using $k$ -Means method on DCT dataset at $Re = 4000$ . ....	59
Figure 6.14. Evaluation metrics for the number of clusters generated using $k$ -Medoids. ....	60
Figure 6.15. Generated clusters using the Hybrid A method at $Re = 4000$ . ....	61
Figure 6.16. Cluster $t$ -SNE distribution using Hybrid A method at $Re = 4000$ .....	62
Figure 6.17. Vortex shedding map using Hybrid A method at $Re = 4000$ . ....	63
Figure 6.18. Generated clusters by Hybrid B method at $Re = 4000$ .....	64
Figure 6.19. Cluster $t$ -SNE distribution using Hybrid B method at $Re = 4000$ .....	65
Figure 6.20. Vortex shedding map using Hybrid B method at $Re = 4000$ . ....	66
Figure 6.21. Evaluation metrics for the number of clusters generated using $k$ -Means. ....	67
Figure 6.22. Generated clusters by Hybrid C method at $Re = 4000$ .....	68
Figure 6.23. Cluster $t$ -SNE distribution using Hybrid C method at $Re = 4000$ .....	68
Figure 6.24. Vortex shedding map using Hybrid C method at $Re = 4000$ . ....	69
Figure 6.25. Overlaid benchmark regimes on vortex shedding map produced with $k$ -Means at $Re = 4000$ . ....	72
Figure 6.26. Overlaid benchmark regimes on vortex shedding map produced with agglomerative at $Re = 4000$ . ....	73
Figure 6.27. Overlaid benchmark regimes on vortex shedding map produced with Hybrid B at $Re = 4000$ . ....	75
Figure 6.28. Force in phase with acceleration, $Cycos\phi$ , contour graph at $Re = 4000$ [7].....	76
Figure 7.1. Dataset sampled nodes in normalized amplitude–wavelength plane [7].....	79
Figure 7.2. Motif extraction for high Reynolds number signals of (a) consistent pattern observed at $(\lambda^*, A^*) = (0.5, 6)$ and (b) a relatively unstable signal observed at $(\lambda^*, A^*) = (0.9, 4)$ .....	80
Figure 7.3. Generated clusters by $k$ -Means method at $Re = 10,000$ .....	81
Figure 7.4. Cluster $t$ -SNE distribution at $Re = 10,000$ using $k$ -Means. ....	82
Figure 7.5. Vortex shedding map using $k$ -Means method at $Re = 10,000$ . ....	83
Figure 7.6. Generated clusters by Agglomerative (complete, cosine) method at $Re = 10,000$ . ....	84
Figure 7.7. Cluster $t$ -SNE distribution using the agglomerative method at $Re = 10,000$ . ....	84

Figure 7.8. Vortex shedding map using the agglomerative method at $Re = 10,000$ .....	85
Figure 7.9. Generated clusters by $k$ -Means method on DCT dataset at $Re = 10,000$ .....	86
Figure 7.10. Cluster $t$ -SNE distribution at $Re = 10,000$ using $k$ -Means on DCT dataset.....	87
Figure 7.11. Vortex shedding map using $k$ -Means method on DCT dataset at $Re = 10,000$ .....	88
Figure 7.12. Evaluation metrics for the number of clusters generated using $k$ -Medoids.....	88
Figure 7.13. Generated clusters using the Hybrid A method at $Re = 10,000$ .....	89
Figure 7.14. Cluster $t$ -SNE distribution using Hybrid A method at $Re = 10,000$ .....	90
Figure 7.15. Vortex shedding map using Hybrid A method at $Re = 10,000$ .....	91
Figure 7.16. Generated clusters by Hybrid B method at $Re = 10,000$ .....	92
Figure 7.17. Cluster $t$ -SNE distribution using Hybrid B method at $Re = 10,000$ .....	92
Figure 7.18. Vortex shedding map using Hybrid B method at $Re = 10,000$ .....	93
Figure 7.19. Evaluation metrics for the number of clusters generated using $k$ -Means.....	94
Figure 7.20. Generated clusters by Hybrid C method at $Re = 10,000$ .....	95
Figure 7.21. Cluster $t$ -SNE distribution using Hybrid C method at $Re = 10,000$ .....	95
Figure 7.22. Vortex shedding map using Hybrid C method at $Re = 10,000$ .....	96
Figure 7.23. Vortex shedding map regions using the Hybrid B method at $Re = 10,000$ .....	98
Figure 7.24. Vortex shedding map regions using the Hybrid C method at $Re = 10,000$ .....	99

# List of Tables

Table 2.1: Flow Regimes for Stationary Smooth Cylinders In Crossflow. ....	7
Table 4.1: Vortex Mode Parameters and Non-dimensional Groups.....	33
Table 5.1: Testing Accuracy of Machine Learning Models .....	44
Table 6.1: Clustering Performance Metrics of $k$ -Means Method at $Re = 4000$ .....	52
Table 6.2: Vortex Shedding Map Cluster Candidates for $k$ -Means at $Re = 4000$ .....	53
Table 6.3: Clustering Performance Metrics of Agglomerative Method at $Re = 4000$ .....	54
Table 6.4: Vortex Shedding Map Cluster Candidates for Agglomerative $Re = 4000$ . ....	56
Table 6.5: Clustering Performance Metrics of DCT dataset using $k$ -Means Method at $Re = 4000$ .....	57
Table 6.6: Vortex Shedding Map Cluster Candidates for $k$ -Means on DCT dataset $Re = 4000$ .....	59
Table 6.7: Clustering Performance Metrics of Hybrid A Method at $Re = 4000$ .....	60
Table 6.8: Vortex Shedding Map Cluster Candidates for Hybrid A at $Re = 4000$ .....	62
Table 6.9: Clustering Performance Metrics of Hybrid B Method at $Re = 4000$ .....	64
Table 6.10: Vortex Shedding Map Cluster Candidates for Hybrid B at $Re = 4000$ .....	65
Table 6.11: Clustering Performance Metrics of Hybrid C Method at $Re = 4000$ .....	67
Table 6.12: Vortex Shedding Map Cluster Candidates for Hybrid C at $Re = 4000$ . ....	69
Table 6.13: Pros and Cons of Ordinary Clustering Methods.....	71
Table 6.14: Comparison of Ordinary Clustering Methods Based on The Silhouette and Dunn Indices .	71
Table 6.15: Pros and Cons of Hybrid Clustering Methods.....	74
Table 6.16: Comparison of Hybrid Clustering Methods Based on The Silhouette and Dunn Indices ...	74
Table 6.17: Final Clustering Performance Metrics of Proposed Methods at $Re = 4000$ .....	76
Table 7.1: Clustering Performance Metrics of $k$ -Means Method at $Re = 10,000$ .....	81
Table 7.2: Vortex Shedding Map Cluster Candidates for $k$ -Means at $Re = 10,000$ .....	82
Table 7.3: Clustering Performance Metrics of Agglomerative Method at $Re = 10,000$ .....	83
Table 7.4: Vortex Shedding Map Cluster Candidates for Agglomerative $Re = 4000$ .....	85
Table 7.5: Clustering Performance Metrics of DCT dataset using $k$ -Means Method at $Re = 10,000$ .....	86
Table 7.6: Vortex Shedding Map Cluster Candidates for $k$ -Mean on DCT dataset $Re = 10,000$ .....	87
Table 7.7: Clustering Performance Metrics of Hybrid A Method at $Re = 10,000$ .....	89
Table 7.8: Vortex Shedding Map Cluster Candidates for Hybrid A at $Re = 10,000$ .....	90
Table 7.9: Clustering Performance Metrics of Hybrid B Method at $Re = 10,000$ .....	91
Table 7.10: Vortex Shedding Map Cluster Candidates for Hybrid B at $Re = 10,000$ .....	93
Table 7.11: Clustering Performance Metrics of Hybrid B Method at $Re = 10,000$ .....	94
Table 7.12: Vortex Shedding Map Cluster Candidates for Hybrid C at $Re = 10,000$ .....	96
Table 7.13: Final Clustering Performance Metrics of Proposed Methods at $Re = 10,000$ .....	97

# Chapter 1

## Introduction

Vortex shedding is an aerodynamic phenomenon when fluid flows around a bluff body, such as a cylinder. Vortex shedding produces unbalanced forces acting on the structure, which causes vortex-induced vibrations (VIV). The majority of research regarding the effects of vortex shedding has been on reducing the unbalanced forces for applications suffering from the phenomenon. VIV arises in many domains, such as the design of skyscrapers, pipelines [1], offshore structures [2], bridges [3], and tube bank heat exchangers [4]. A prominent example of VIV mitigation is the helical strakes added to the exterior of tall smokestacks or chimneys, as shown in Figure 1.1 [5].



Figure 1.1. Helical strakes on smoke towers for the mitigation of VIV [6].

Bladeless wind turbines are a new concept of wind harvesting machines that utilize VIV to oscillate a vertical cylinder. Instead of reducing the oscillations, the main principle of bladeless wind turbines is to take advantage of bluff bodies' natural phenomenon to extract renewable energy from the motion.

The generation and shedding of large coherent vortex structures occur in distinct modes. The most common modes are 2S, described by two single opposingly spinning vortices shed per oscillating period, and 2P, described by a pair of opposing spinning vortices shed per oscillating period. The hydrodynamic signatures of the unsteady wake behind the bluff body depend on the body's intrinsic properties and movement.

Forced vibration experiments are often used to study vortex shedding behaviour and rely on consistent use of the non-dimensional parameters of the prescribed motion. Morse and Williamson [7]

produced a vortex shedding map for the normalized amplitude–wavelength plane by conducting numerous forced vibration experiments at a Reynolds number of 4000. The authors conducted 5860 experimental runs in a water channel to obtain high-resolution vortex shedding regimes, as shown in Figure 1.2.

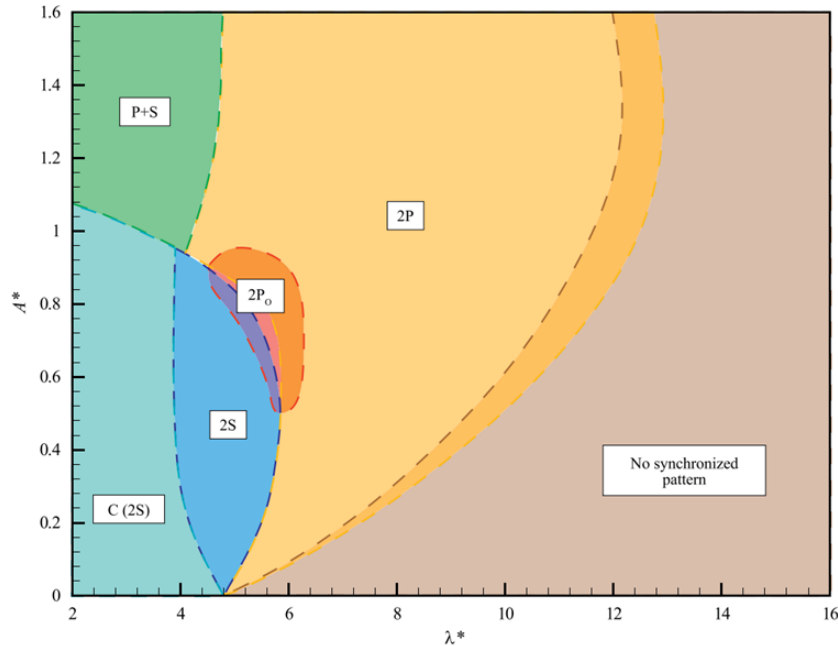


Figure 1.2. Forced vibration vortex shedding map in normalized amplitude–wavelength space [7].

## 1.1 Objectives

The main objective of the thesis was to develop a data-driven approach to generating vortex shedding maps for a cylinder under prescribed motion. This larger objective was discretized into three smaller goals: namely,

1. Validate the use of local flow measurements for vortex shedding identification.
2. Develop data-driven methods and validate their performance in the vortex shedding map using a reference case at a low Reynolds Number.
3. Extend the method from Objective 2. to generate a vortex shedding map at a higher Reynolds number case.

A quantitative comparison of various machine learning models trained using differing simulated vortex shedding local flow measurements and corresponding feature noise analysis provided an effective wake classification strategy to satisfy the first objective. Next, Objective 2 was satisfied by developing an unsupervised clustering method to extract a low number of well-defined clusters that are rooted in the flow physics of the low Reynolds number case to reproduce the benchmark vortex shedding map. Finally, the final objective was met by generating vortex shedding maps using the methods from Objective 2. for higher Reynolds numbers and discussing the insights gained on the underlying dynamical regimes of the physical system.

## 1.2 Contributions

The effort of this dissertation deviates from previous work concerning vortex shedding mode classification by demonstrating the gains of representing local fluid measurements in the frequency domain for the application of oscillating cylinders. Also, the frequency domain feature vectors method was expanded for higher complexity vortex modes. The results highlight the ability of the proposed schema to accurately identify vortex structures in the wake of an oscillating cylinder with reduced input data and computational resource required, which provides validation for the unsupervised clustering of local flow measurements.

Previous unsupervised clustering approaches have shown promising results for identifying and dissection of vortex shedding modes. However, the methods require extensive input data to compute the entire flow field. This dissertation's effort addresses the opportunity to develop a data-driven method using unsupervised clustering of vortex shedding modes from local flow signatures sampled in the wake. Furthermore, the primary contribution of this method is its use in generating vortex shedding maps for high Reynolds numbers, which have been limited in literature due to the increased computational cost and complex dynamics. The unsupervised clustering approach used in this effort varies from previous studies by clustering local flow measurements and leveraging the benefits of other flow field sensors.

## 1.3 Thesis Organization

The structure of this dissertation includes an introduction, a background, literature review, methodology and the remaining chapters are divided into three sections: 1) Mode Classification using Machine Learning (Chapter 5), 2) Vortex Shedding Map Generation at Low Reynolds Number (Chapter 6), and 3) Vortex shedding map Generation at High Reynolds Number (Chapter 7).

The contents of these chapters are summarized as follows:

- Chapter 5 presents an effective wake classification strategy, applying machine learning models trained using fluid sensor data. The demonstrated ability to classify vortex shedding modes using the local flow measurement dataset structure is a pivotal proof-of-concept for applying the following clustering analysis.
- Chapter 6 details the method for generating vortex shedding maps using unsupervised clustering. The results of the vortex shedding map generation method are compared to a pre-existing map to validate the method's reproducibility for its' application to unknown regimes.
- Chapter 7 extends the vortex shedding map generation method to an unknown map domain and quantifies the performance of the clustering method for more complex flow regimes.

# Chapter 2

## Background

### 2.1 Introduction

This chapter presents an overview of the fundamental concepts of vortex-induced vibration (VIV) due to vortex shedding and time series clustering. Vortex-induced vibration is first introduced through the theoretical analysis and the subsequent free and forced experimental vibration setups used to study the vortex shedding modes. Next, time series clustering is discussed, emphasizing subsequence clustering, which is the focus of this thesis. The aspects of the clustering analysis are presented, including explaining the roles of time series representation, similarity measures, clustering algorithms, and evaluation metrics on the analysis.

### 2.2 Vortex-Induced Vibration

The phenomenon of vortex-induced vibrations due to vortex shedding of bluff bodies is introduced first with a theoretical analysis of a fixed cylinder. Next, the two primary methodologies used to study VIV are introduced, focusing on forced vibration and its implications in investigating vortex shedding modes at varying Reynold numbers.

#### 2.2.1 Theory

The theoretical analysis for a simplified two-dimensional fixed cylinder in uniform flow allows for a preliminary understanding of vortex-induced vibrations. The vortex-induced vibration for a cylinder of uniform diameter,  $D$ , in crossflow,  $U$ , with transverse oscillation motion is illustrated in Figure 2.1.

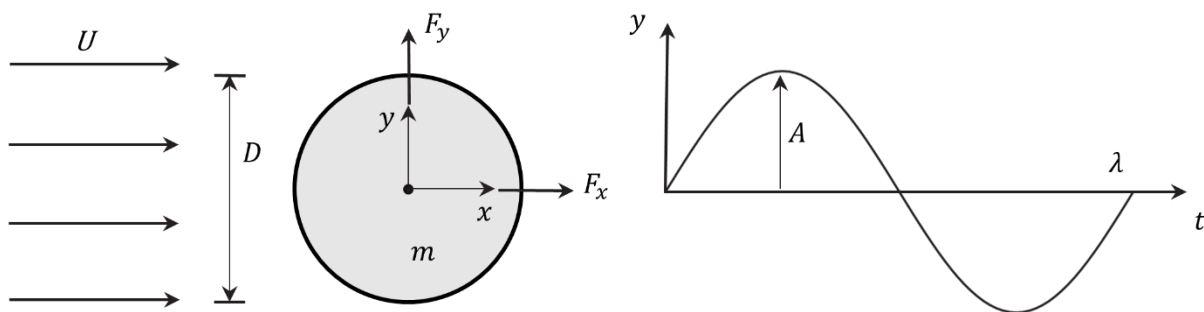


Figure 2.1. Illustration of a cylinder in free stream and sinusoidal vertical motion.

The parameters of interest for vortex-induced vibration include the Reynolds number ( $Re$ ), oscillation amplitude ( $A$ ), oscillation wavelength ( $\lambda$ ), and oscillation frequency ( $f$ ). The Reynolds number for the case of a cylinder in crossflow is defined in Equation (2.1).

$$Re = \frac{\rho U D}{\mu} \quad (2.1)$$

The Reynolds number is a function of the fluid density,  $\rho$ , and fluid viscosity,  $\mu$ . The oscillation amplitude is normalized using the diameter of the cylinder to obtain a non-dimensional amplitude of the sinusoidal motion,  $A^*$ , described in Equation (2.2).

$$A^* = \frac{A}{D} \quad (2.2)$$

The oscillation wavelength is converted to the non-dimensional parameter of the normalized wavelength,  $\lambda^*$ , using the cylinder's diameter and the frequency representation defined in Equation (2.3).

$$\lambda^* = \frac{\lambda}{D} = \frac{U}{f D} \quad (2.3)$$

Under sinusoidal forcing, the motion of the cylinder in the vertical axis is approximated by a sinusoidal function represented by,

$$y(t) = A \sin(\omega_s t - \phi) \quad (2.4)$$

where  $\omega_s = 2\pi f_s$  is the circular vortex shedding frequency as a function of the dominant vortex shedding frequency,  $f_s$ , and the phase shift of the oscillation,  $\phi$ . The dominant vortex shedding frequency,  $f_s$ , is defined based on its relationship to the free stream velocity and diameter as

$$f_s = St \frac{U}{D} \quad (2.5)$$

where  $St$  is the Strouhal number [8]. The Strouhal number represents the ratio of a characteristic flow time to a characteristic oscillation time. The force acting on the cylinder from the vortex shedding is quantified by a lift force that acts traverse to the flow direction. The lift force per unit length is represented by the sinusoidal function in Equation (2.6).

$$F_L(t) = \frac{1}{2} \rho U^2 D C_L \sin(\omega_s t - \phi) \quad (2.6)$$

The lift force can be redefined in dimensionless form using the time-varying coefficient of lift by rearranging Equation (2.6).

$$C_L(t) = \frac{F_L(t)}{1/2 \rho U^2 D} \quad (2.7)$$

The drag force acting on the cylinder is another useful parameter in the vortex shedding effect, primarily for two-dimensional vibration analysis. The drag coefficient is defined by Equation (2.8)

$$C_D(t) = \frac{F_D(t)}{1/2 \rho U^2 D} \quad (2.8)$$

### 2.2.1.1 Effect of Reynolds Number

The Reynolds number directly affects the vortex shedding behaviour of a fixed cylinder. Vortex shedding can be analyzed from the perspective of the Strouhal number and its relationship with the Reynolds



Number. The Strouhal number for a stationary two-dimensional cylinder with varying Reynolds numbers is shown in Figure 2.2.

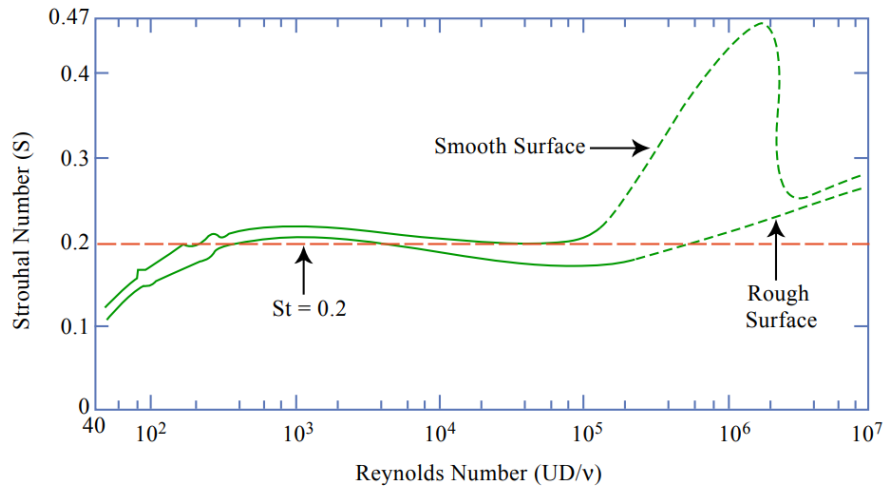


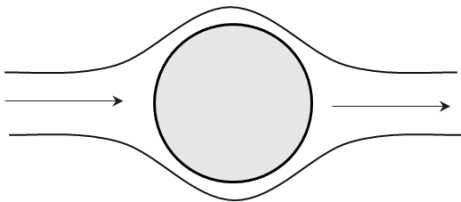
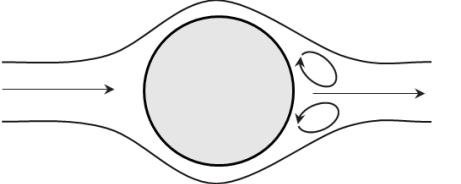
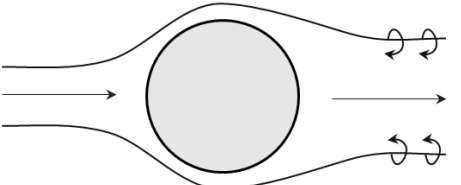
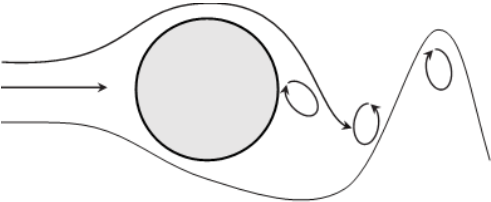
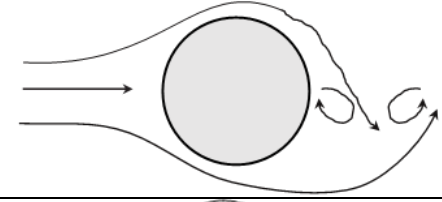
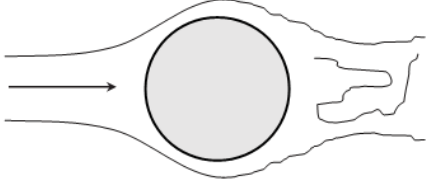
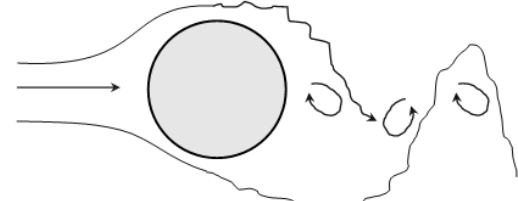
Figure 2.2. Strouhal Number and Reynolds Number relationship [9].

The following regimes are outlined by analyzing the changes of Strouhal number and the corresponding vortex shedding behaviour:

1. No vortex shedding is present for minimal Reynolds numbers,  $Re \lesssim 40$ , and the flow patterns vary from creeping flow to recirculation in the wake. At this low Reynolds number, the Strouhal number is equal to zero.
2. The regime between Reynolds number  $40 \lesssim Re \lesssim 300$ , vortex shedding occurs with Laminar Vortex Streets and increasing Strouhal number and lift coefficient.
3. For the range of Reynolds number  $300 \lesssim Re \lesssim 250,000$  vortex shedding persists and contains several dynamic changes while the Strouhal number stays relatively constant  $St \approx 0.2$ .
4. The regime between Reynolds number  $250,000 \lesssim Re \lesssim 500,000$ , vortex shedding disappears for smooth cylinders with low turbulence stream.
5. The final domain is associated with a high Reynolds number,  $Re \gtrsim 10^6$ , where fully turbulent vortex shedding occurs at a higher Strouhal number  $St \approx 0.26$ .

The regimes distinguished have the approximate flow fields summarized in Table 2.1.

Table 2.1: Flow Regimes for Stationary Smooth Cylinders in Crossflow

Flow Illustration	Description	Reynolds/Strouhal Number
	Unseparated Flow.	
	Recirculation in the wake.	$Re \lesssim 40$ $St = 0$
	Braid vortices are roughly parallel to the incoming flow [10].	
	Laminar vortex shedding.	$40 \lesssim Re \lesssim 300$ $St \propto Re$
	Turbulence on vortex shedding.	$300 \lesssim Re \lesssim 250,000$ $St \approx 0.2$
	Turbulent boundary layer transition, no shedding behaviour.	$250,000 \lesssim Re \lesssim 500,000$
	Re-established turbulent vortex shedding.	$Re \gtrsim 10^6$ $St \approx 0.26$

Two regions have been proposed to study vortex shedding divided by a critical Reynolds Number [10]. The first region, regarded as subcritical, is typically below  $Re \approx 1000$ , which includes the first regime of vortex shedding onset by instabilities at Reynolds number  $Re \approx 40$  [10]. The subsequent instability occurs around  $Re \approx 300$  which introduces turbulence in the wake. The next region, regarded as the critical regime, begins at approximately  $Re \approx 250,000$ , where the boundary layer transitions from laminar to turbulent, affecting the flow separation points and the process of vortex formation [10]. A significant drop in drag coefficient marks the onset of the critical regime, referenced as the drag crisis, which increases as the Reynolds number increases over the regime [10]. Above this regime is the supercritical regime, which begins at approximately  $Re \approx 10^6$  which includes the re-establishment of the turbulent vortex shedding.

These outlined regimes are approximations of the two-dimensional flow dynamics in the wake of a stationary cylinder, and the onset of each regime is sensitive to a variety of factors such as surface roughness and turbulence intensity of the free stream. The Reynolds number has similar effects on a cylinder free to oscillate in the transverse direction. Specifically, the Reynolds number affects the boundary layer's development and the separation points for the moving cylinder.

### 2.2.2 Free and Forced Vibration

There are two primary methodologies to study vortex-induced vibration for a mounted cylinder of circular diameter in uniform crossflow: namely, free and forced vibration. A free vibration experiment of VIV allows the cylinder to oscillate due to the external and unbalanced forces produced by the fluid [10]. The structure will vibrate corresponding to the vortex shedding mode, oscillating amplitude, and frequency when under synchronization conditions. The oscillating cylinder for free vibration is modelling by mounting the cylinder on linear springs with constant  $k$ , a linear damper with constant  $b$ , and cylinder mass,  $m$ , as shown in Figure 2.3.

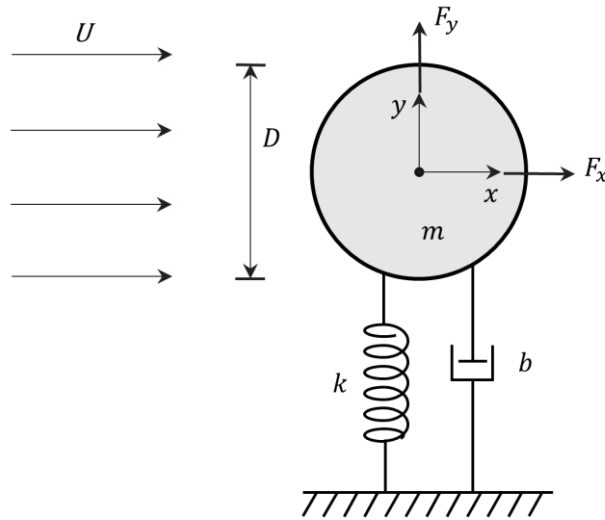


Figure 2.3. Modelled cylinder for transverse oscillations in crossflow.

The equation of motion for the vibrations of an elastically mounted cylinder is defined by Equation (2.9).

$$m \frac{d^2}{dt^2} y(t) + b \frac{d}{dt} y(t) + k \cdot y(t) = F(t) \quad (2.9)$$

The fluid force is denoted by  $F(t)$  is comprised of two parts: namely, the component in phase with velocity and the component in phase with the acceleration [10].

Forced vibrations models, which are considered in this study, are used to approximate the free vibration model described by prescribing the oscillatory motion of the cylinder. Forced vibration models are used in experiments to better control the oscillations through resonance and dampening. The studies using forced vibration are often conducted in the parameter space of non-dimensional frequency and amplitude to standardize the motion of frequency,  $f$ , amplitude,  $A$ , and incoming flow velocity,  $U$  [11].

### 2.2.2.1 Wake Excitation Regions and Lock-in

Free vibrations occur in the wake excitation (positive energy) regime, where the vibration energy is only transferred from the fluid energy marked by positive lift coefficients in phase with the velocity [10]. In forced vibration, the wake capture region is present where the vortices match the frequency of the cylinder vibration as opposed to the Strouhal frequency. The positive energy and wake capture regions are shown in Figure 2.4.

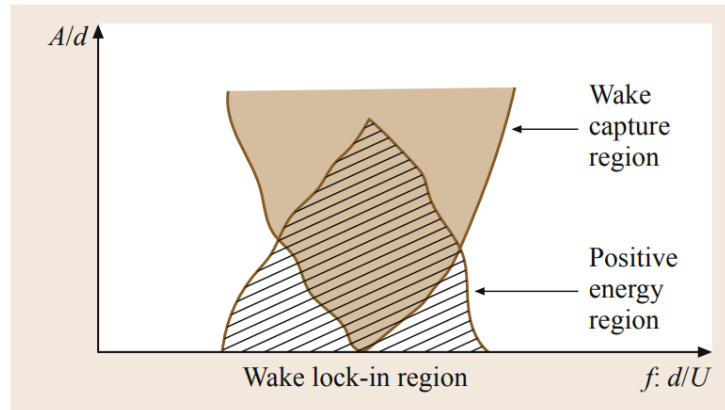


Figure 2.4. Free vibration and forced vibration in fluid-structure energy transfer regions. [10].

The overlap of the wake capture and wake excitation regions is called the lock-in region. The lock-in region is a state of resonance.

### 2.2.2.2 Forced Vibration for the Study of Free Vibration

The forced vibration experiments provide insights into free vibration vortex shedding modes by exploring the parameter space to match flow-induced vibration patterns. Previous studies discovered limitations of forced vibration experiments that observed negative fluid excitation where free vibration was expected to occur [12]. Morse and Williamson [13] addressed the debate of the validity of the predicting power of forced vibration by finding that thorough matching of amplitude, frequency, and Reynolds number resulted in consistent fluid forces. The forced vibration model used in this dissertation was controlled with respect to matching the non-dimensional amplitude and frequency with special consideration of Reynolds number to justify the use of this model.

### 2.2.3 Vortex Shedding Modes

The generation and shedding of large coherent vortex structures occur in distinct modes. The vortex shedding patterns produced by two-dimensional investigations will be the focus of the following review.

Williamson and Roshko [14] built upon the fundamental study of Bishop and Hassan [15] to explore the vortex shedding patterns from a forced oscillation of a cylinder. Williamson and Roshko [14] studied the parameter space by varying the Reynolds number from 300 to 1000 to generate a map of the synchronized patterns observed. The vortex shedding modes identified near the lock-in regions are denoted as C(2S), 2S, 2P, and P+S, as shown in Figure 2.6. The mode 2S is described by two single opposingly spinning vortices shed per oscillating period, and 2P, characterized by a pair of opposing spinning vortices shed per oscillating period. The mode C(2S) is similar to the 2S vortex structure as smaller vortices coalesce in the near field to produce the larger structures in the far-field. A P+S mode is an asymmetric form of the 2P mode described by a pair of vortices and a single vortex shed per oscillating period. An additional vortex shedding mode denoted as 2P+2S comprises two pairs of opposingly spinning vortices separated by two single vortices being shed. The boundary between P+S and 2P was well established for Reynolds number 300 - 1000, but the sensitivity to Reynolds number of these modes produced the P+S mode in the 2P region for Reynolds number  $Re < 300$ .

Morse and Williamson [7] expanded the work Williamson and Roshko [14] to  $Re = 4000$  to produce an extensive map of the vortex shedding modes. The authors conducted 5680 experimental runs in the parameter space to investigate areas of interest. The existence of the C(2S), 2S, 2P, and P+S modes was confirmed in the regions expected. The authors identified a new mode at the transition boundary of the 2S and 2P modes named '2P Overlap' or reduced '2PO'. The 2PO mode is described by two pairs of vortices being shed per cycle with one vortex in each oscillation much weaker and intermittently switches between the 2S and 2P modes. The weaker secondary vortex decays rapidly as the vortices travel downstream, resembling 2S. The fluid excitation represented by force in phase with velocity determines which modes should be expected for free vibration. The fluid excitation of the P+S mode was measured to be strongly negative, concluding that this mode would not appear for free vibration cases. The boundaries of the vortex shedding modes were outlined in the normalized amplitude-wavelength plane, as shown in Figure 2.5.

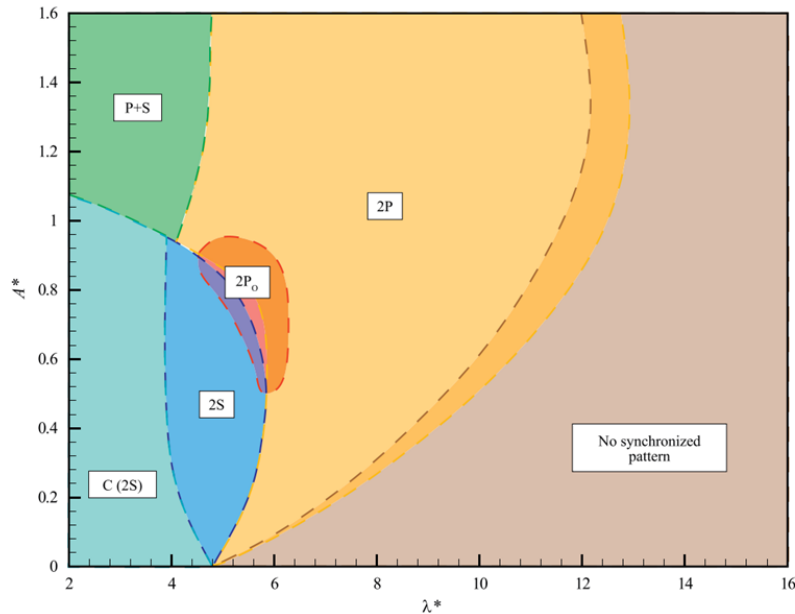


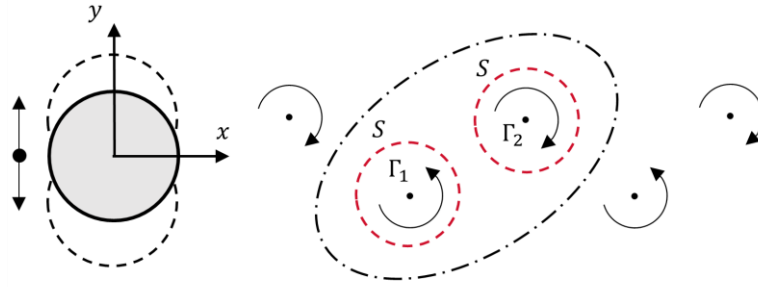
Figure 2.5. Vortex shedding map in the normalized amplitude–wavelength plane for forced vibration [7].

## 2.2.4 Vortex Shedding Modes at High Reynolds Number

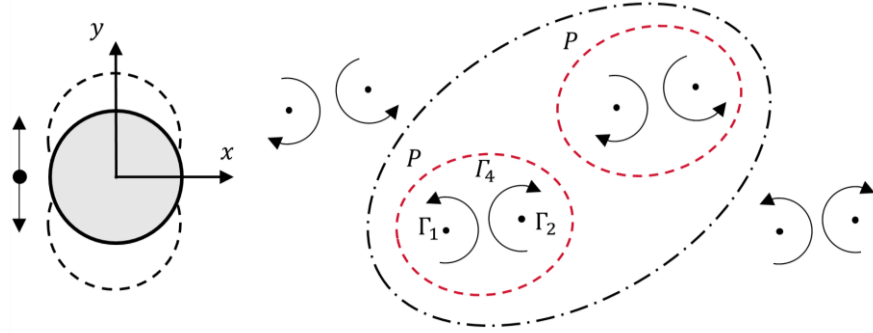
Forced vibration analysis has been implemented for high Reynolds cases. Wu et al. [16] studied vortex shedding patterns between Reynolds number 35,000 - 130,000 using computational fluid dynamic simulations. The authors observed the first transition at approximately  $Re \approx 4000$  where the 2S and 2P mode was identified. A transitional mode between 2S and 2P was identified and named by the authors quasi-2P which aligns with the 2PO identified by Morse and Williamson [7]. At higher Reynolds numbers, 2S patterns were observed at  $Re = 35,000$  and 2P at  $Re = 50,000$  and  $55,000$ . A 2P mode was observed in most cycles at  $Re = 65,000$  with occasional 1P+2S patterns. A P+S mode was observed at  $Re = 70,000$  but was only shed on one side and occasional on both sides but not defined enough for mode classification. A relatively stable mode identified as 2P+4S was observed at  $Re = 78,000$ . The mode is described by a pair of vortices and two single vortices being shed, in that order, on each side of the oscillation. A similar 2P + 4S mode was observed at  $Re = 110,000$  and  $120,000$ , but the order of shedding differed, a single vortex followed by a pair of vortices was shed then the final single vortex was shed. The authors identified a final highly regular mode at  $Re = 130,000$  denoted 2P+8S.

The work by Wu et al. [16] was extended by Zhang et al. [17] to study the effect of turbulence intensity on the vortex shedding modes. The authors utilized the same CFD methodology for the simulation of forced vibration but experimented with three levels of free-stream turbulence intensity (0.2%, 1%, 5%). The termed quasi-2P vortex shedding mode, or 2PO in the terminology of Morse and Williamson [7], was observed at  $Re = 30,000$  across all turbulence intensities though the strength and cohesiveness decreased for greater intensities. The vortex shedding patterns begin to differ between turbulence intensities at the Reynolds number 50,000. Two vortex shedding patterns was observed for the lowest turbulence intensity case, P+QP and T+QT, where "Q" represents the same quasi-state such that one of the vortices is weaker than the rest. At turbulence intensity of 1%, the P+QP mode transforms into a P+S mode due to the increased diffusion, which causes the weaker vortex in the QP portion to cancel with the stronger vortex to create a single S vortex. At Reynolds number  $Re = 70,000$ , a vortex structure denoted T+S+P is observed clearly at turbulence intensity 0.2%, which dissipates to a T+P mode at intensity 1%. Similarly, to the  $Re = 50,000$  case, no uniform vortex shedding behaviour was observed for the highest turbulence intensity 5%. For the last case at Reynold number  $Re = 100,000$ , the vortex mode T+S+S+T was observed at turbulence intensity 0.2% and a T+QT+S mode for turbulence intensity 1%. At the turbulence intensity of 5%, the vortex structures become elongated due to the high flow velocity of the free steam and significantly mixed due to the high turbulence energy. The general conclusion taken from the study is that when the Reynolds number increases, the number of vortices shed each oscillation increases. Furthermore, the turbulence intensity of the incoming stream has a dissipation effect that causes the vortices to become weaker and increasingly difficult to distinguish.

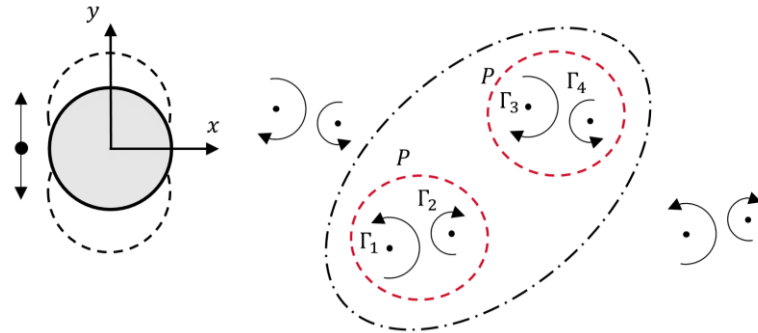
The idealized point vortex representation of the standard wake modes, where the strength of the vortex is denoted  $\Gamma$ , is shown in Figure 2.6.



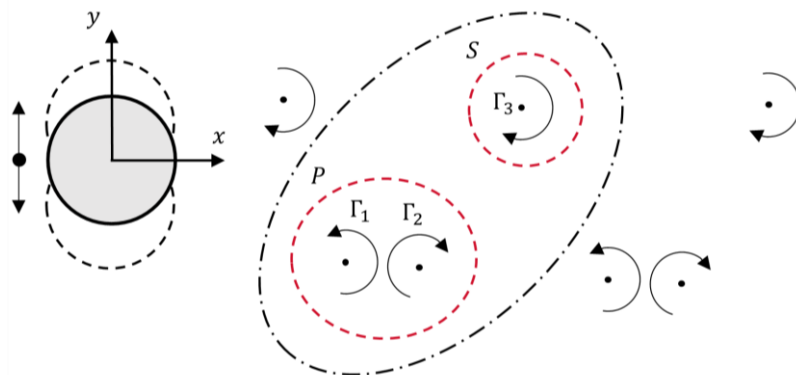
(a) Ideal 2S wake mode.



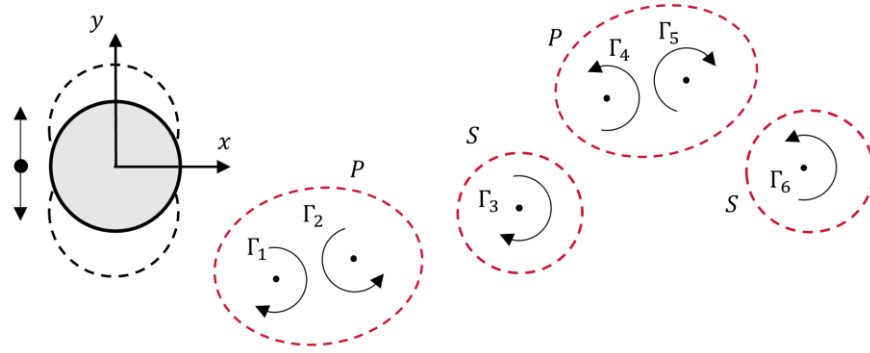
(b) Ideal 2P wake mode.



(c) Ideal 2PO wake mode, where  $\Gamma_2, \Gamma_4 < \Gamma_1, \Gamma_3$ .



(d) Ideal P+S wake mode.



(e) Ideal 2P+2S wake mode.

Figure 2.6. Point vortex models for the wake regimes of (a) 2S, (b) 2P, and (c) 2PO, (d) P+S, (e) 2P+2S.

### 2.3 Time Series Clustering

Machine learning is the process of leveraging statistical models trained on data to answer questions and provide insight for data mining tasks. Machine learning tasks are organized into three main categories depending on the inputs: namely, supervised, unsupervised, and reinforcement learning. Unsupervised learning uses the input data structure to provide insights without referencing labels associated with the data. Unsupervised learning is often referred to as data mining, and if the input data is assumed to have discrete structures, the task is referred to as clustering.

Clustering analysis is a subset of unsupervised machine learning in which a dataset is decomposed into groups, or "clusters," based on detected shared structures in the data. The unsupervised nature of clustering means there is limited external information on the data structure other than the intrinsic properties. Clustering analysis applied to time series data has recently increased importance due to the onset of cloud computing and big data capable of storing large amounts of data in the fields of environmental science, medicine, finance, engineering, and politics. Clustering and unsupervised machine learning have gained prevalence due to the cost of labelling. The performance of supervised machine learning models depends significantly on the quality of the labels provided, and the cost of acquiring the labels and cleaning incorrectly labelled samples comes at a high cost, making supervised methods unpractical.

A time series is a sequence of data points indexed at specific points in time denoted,  $X_t = \{X(t)\}, t \in T$ , where the observation at the time,  $t$ , is a subset of allowed timesteps,  $T$ . The literature's time series clustering approach is discretized into three main categories: whole time series, subsequence, and time point clustering.

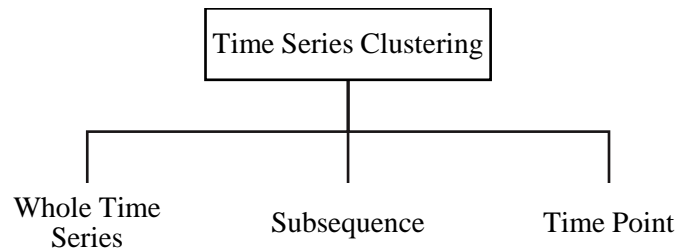


Figure 2.7. Taxonomy of time series clustering approaches.



Whole-time series clustering utilizes the entire series length as the clustering object. The approach of whole-time series clustering often relies on dimensionality reduction due to the complexity of distance metrics [18]. Subsequence time series clustering extracts shorter sequences from the entire time series as the objects for clustering. Sequences can be extracted using a sliding window or novel data mining techniques such as shapelet/motif discovery [19]. Timepoint clustering incorporates the temporal proximity of the time series data with the values as the clustering objects [18].

### 2.3.1 Subsequence Time Series Clustering

Subsequence time series clustering is clustering frequent patterns within a more extended time series. Shapelet/motif discovery is the process of identifying frequent patterns within a time series and has been actively researched in the data mining community. The nomenclature of motif discovery is often used in the computational biology application of shapelet discovery [20]. The term motif originates from finding DNA motifs that are nucleic acid sequence patterns with biological significance [21]. The application of pattern discovery has extended beyond the biological field to severe weather prediction [22], wind generation [23], face image recognition [24], motion graphs [25], and electrocardiogram (ECG) anomaly detection [26]. An example of a motif within a time series is shown in Figure 2.8.

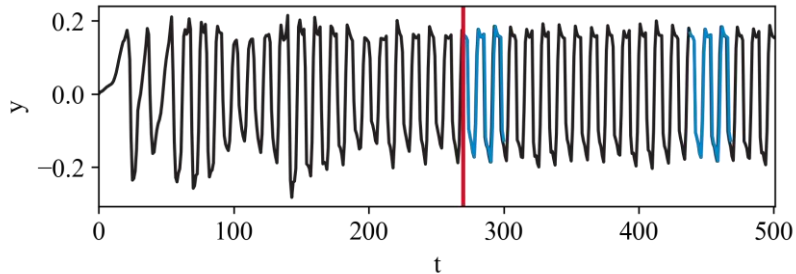


Figure 2.8. Example of a motif discovered in original time series.

Subsequence time series clustering is suitable for the application of classifying vortex shedding modes since the specific patterns can be identified from full-time series where there may be non-steady-state behaviour of the sign. Furthermore, multiple shedding modes may be present in a time series, and sub sequencing isolates the wanted behaviour.

In the field of subsequence time series clustering, Keogh and Lin [27] published a pivotal study that claimed that clustering of time-series subsequences is meaningless. Specifically, the subsequence clusters extracted by any clustering algorithm are essentially random. The authors demonstrated that for subsequences extracted using a sliding window, the global mean for the clusters is a straight line. This fact implies that the weighted average of  $k$  clusters must sum to a straight line as well. The linear constraint on the dataset is not trivial for all datasets, invalidating most subsequence clustering research. Depending on the dataset, extracting subsequences using a sliding window can yield many trivial matches that produce a dense clustering subspace.

The importance of motif/pattern discovery in the application of subsequence clustering is apparent to overcome the constraints of sliding window extraction. Motif extraction addresses the meaningless claim by explicitly disregarding trivial matches in the mining procedure and defining a subset of the data instead of the entire dataset. Chen [28] investigated the meaningless claim and attributed the problems to using Euclidean distances for clustering shape-based similarities. Special consideration is required for

using Euclidean distances in clustering analysis or the use of shape-based distance metrics such as Dynamic Time Warping.

The data mining field has developed numerous time series motif discovery algorithms which vary in methodology based on the designed application. Torkamani and Lohweg [29] surveyed time series motif discovery algorithms. The algorithms for motif discovery depend on the algorithmic exactness, low dimensional representation, and the similarity measure [30].

### 2.3.2 Cluster Analysis

The cluster analysis of time series clustering contains four main components in the process: namely, the representation of the time-series data, similarity/dissimilarity measures, clustering algorithms, and evaluation metrics.

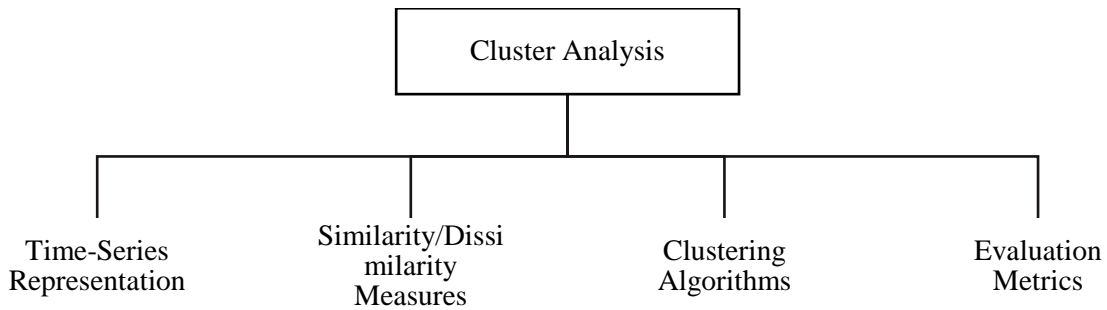


Figure 2.9. Taxonomy of cluster analysis.

### 2.3.3 Time Series Representation

The effectiveness of the cluster analysis greatly relies on the representation of the time series objects. The entire time series object can be utilized for the clustering task, but dimensionality reduction can achieve performance gains.

Dimensionality reduction is essential in machine learning pre-processing because it decreases time and space complexity offering more detailed exploration and visualization, and produces a simpler model. A simpler clustering model offers many benefits, such as improved generalizability to unseen instances, robustness to noise and outliers, and reduced training samples. Time series representation through dimensionality reduction is often used for the time series clustering category of whole-time series but is not restricted [18]. Using representation methods for time series is to characterize the data in a lower dimension through feature extraction that preserves the global structure. Time series representation transforms a time series  $X_t = \{x_1, x_2, \dots, x_t, \dots, x_T\}$  into a lower dimension  $X'_t = \{x'_1, x'_2, \dots, x'_t, \dots, x'_{T'}\}$  where  $T' < T$ .

Reducing the dimensionality of the time series can benefit clustering analysis by reducing the memory costs and speeding up clustering by reducing computational cost for distance calculation [18]. There are four basic categories of time series representation: namely, data-adaptive, non-data adaptive, model-based, and data-dictated. Data adaptive representation selects a standard representation for all the instances in the dataset based on the minimum global reconstruction error [31]. Data-adaptive methods include Symbolic Aggregate Approximation (SAX) [32] and Piecewise Linear Approximation (PLA). Conversely, the non-data adaptive approach constructs an approximate representation based on the local properties of the dataset. Examples of non-data adaptive methods include Discrete Fourier Transform

(DFT), Discrete Wavelet Transform (DWT), and Discrete Cosine Transform (DCT). Model-based methods represent the time series as the parameters of an underlying model that produced the sets. Data-dictated representation methods automatically determine the compression ratio of the raw time series. The Clipped method is an example of a data-dictated method. The non-data adaptive example of DCT for time series representation methods is presented for this application due to their heritage in natural and stationary signals [18].

### 2.3.3.1 Discrete Fourier and Cosine Transforms

The discrete Fourier transform is derived using a Fourier analysis which expresses a signal as a summation of the frequency spectrum components. An example of the decomposition of a signal in both the time and frequency domain is shown in Figure 2.10.

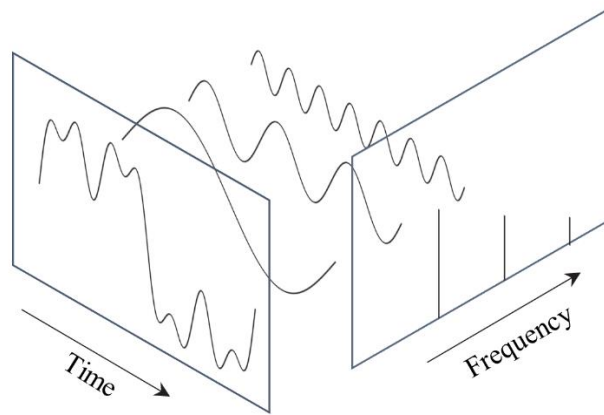


Figure 2.10. Signal representation and decomposition in the time and frequency domain.

The discrete Fourier transform is calculated by using the discrete parts of the transform. The DFT is a non-data adaptive method and uses spectral analysis. The algorithm can be quickly determined using the fast Fourier transform that computes the matrix in  $O(n \log n)$  time [33]. The discrete cosine transform is similar to the discrete Fourier transform in the sense that it represents the time series in its components but instead uses the sum of the cosine terms only.

### 2.3.4 Similarity Measure

Traditional clustering approaches rely on quantifying the similarity or dissimilarity between the time series to combine into similar clusters. The following section will discuss similarity measures on the application of univariate time series. The objectives of the distance measures are subdivided into three main sections.

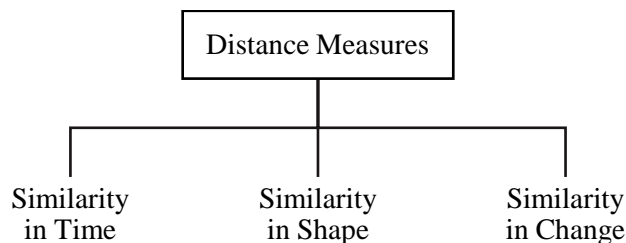


Figure 2.11. Taxonomy of distance measures.

### 2.3.4.1.1 Similarity in Time

The similarity in time distance objectives calculates the similarity based on the time steps relating to time series that are highly correlated. The similarity in time measures is also known as lock-step metrics since the distance is calculated by comparing the  $i$ -th point of one time series to the  $i$ -th point of another series.

The similarity in time distances can be computationally expensive for large raw time series since the distance is calculated for each time step. An example of similarity in time distance concerning time series is the Minkowski distance and its special cases of Euclidean and Manhattan distances.

#### **Minkowski Distance**

The Minkowski distance is a generalized distance metric which, considering two time series,  $X_{iT} = \{x_{i1}, x_{i2}, \dots, x_{iT}\}$ , and  $X_{jT} = \{x_{j1}, x_{j2}, \dots, x_{jT}\}$ , is calculated by Equation (2.10).

$$d(X_{iT}, X_{jT}) = \left( \sum_{k=1}^T |x_{ik} - x_{jk}|^p \right)^{1/p} \quad (2.10)$$

The variable  $p$  can be substituted to generate varying distance metrics such as Manhattan ( $p = 1$ ) and Euclidean ( $p = 2$ ).

#### **Manhattan (City block) Distance**

The Manhattan distance or city block attributes its namesake from calculating the absolute distance between points or the distance between blocks in a city [34]. The Manhattan distance between the time series  $X_{iT}$  and  $X_{jT}$  is defined by Equation (2.11) generated by substituting  $p = 1$  into Equation (2.10).

$$d(X_{iT}, X_{jT}) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{iT} - x_{jT}| \quad (2.11)$$

#### **Euclidean Distance**

The most popular distance measure is the Euclidean distance, often referred to as a one-to-one distance. The meaning of one-to-one refers to the calculation of the distance of each point in sequential order. The Euclidean distance is calculated by substituting  $p = 2$  into Equation (2.10), yielding Equation (2.12)[34].

$$d(X_{iT}, X_{jT}) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{iT} - x_{jT})^2} \quad (2.12)$$

The Euclidean and Manhattan distance share the mathematical properties of being non-negative and possessing the identity of indiscernible. The identity means the distance to an object to itself will be zero. Furthermore, both measures are symmetric, implying that the distance will be the same regardless of the order.

The advantages of the Minkowski-based metrics are the linear complexity and the ease of implementation. The use of these metrics, and specifically Euclidean distance, are competitive in clustering but have limitations to exposure to noise and misalignments in time.

#### **Correlation**

Correlation-based distances are another type of similarity in time metric and consider the time series object as a vector to compute the distance. The correlation implementation for two time series,  $u$  and  $v$ , follows the vector notation

$$d(u, v) = 1 - \frac{(u - \bar{u}) \cdot (v - \bar{v})}{|u - \bar{u}| |v - \bar{v}|} \quad (2.13)$$

where  $\bar{u}$  and  $\bar{v}$  are the mean of the vectors. The  $|x|$  denotes the Euclidean length of the vector.

### Cosine Similarity

The cosine similarity measures the similarity of the vectors by effectively calculating the angle between the vectors.

$$d(u, v) = \frac{u \cdot v}{|u||v|} \quad (2.14)$$

Since the cosine measure quantifies the orientation of the vectors, it is useful when the magnitude and weights are trivial such as when the time series objects are represented as frequencies.

#### 2.3.4.1.2 Similarity in Shape

Distance measures that quantify similarity in shape will find similar patterns of change regardless of time points. The similarity in shape measures is also known as elastic metrics since the distance can be calculated by comparing a single point to many other points or to no other points. An example of a similarity in shape metric is the elastic method of Dynamic Time Warping (DTW).

### Dynamic Time Warping

Dynamic time warping (DTW) is an elastic measure that addresses the limitations of one-to-one metrics of the similarity in time category. Dynamic time warping determines a warping path of the time axis between the time series to achieve the best alignment, minimizing the distance. An  $n \times m$  matrix is constructed between the time series of lengths  $n$  and  $m$ ,  $X_i = \{x_{i1}, x_{i2}, \dots, x_{in}\}$ , and  $X_j = \{x_{j1}, x_{j2}, \dots, x_{jm}\}$ , of the distances between each point, often in Euclidean distance.

$$distMatrix(X_i, X_j) = \begin{bmatrix} d(x_{i1}, x_{j1}) & \cdots & d(x_{i1}, x_{jm}) \\ \vdots & \ddots & \vdots \\ d(x_{in}, x_{j1}) & \cdots & d(x_{in}, x_{jm}) \end{bmatrix}$$

The objective of dynamic time warping is to find the warping path,  $W = \{w_1, w_2, \dots, w_k\}$ , where  $\max(m, n) \leq k \leq m + n - 1$ , which minimizes the distance between the time series objects [35]. The dynamic time warping distance is defined as the associated continuous element path that minimizes the function in equation (2.15).

$$d(X_i, X_j) = \min \sqrt{\sum_{k=1}^K w_k} \quad (2.15)$$

The computation of the distance matrix and the warping path is expensive and often uses dynamic programming for its computation [18]. Dynamic time warping has the advantage over the similarity in time measures of handling objects of varying lengths and time series out of phase but still have similar shapes.

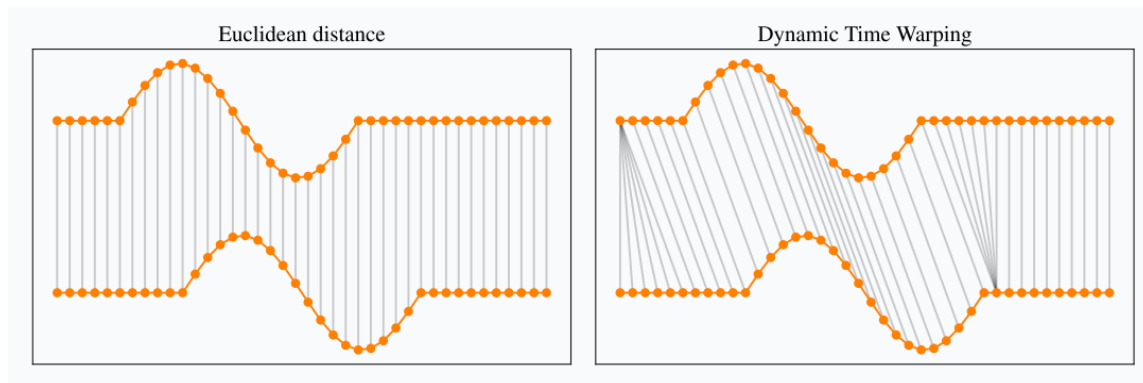


Figure 2.12. Differences in distance measurements between Euclidean and Dynamic Time Warping [36].

The selection of suitable distance metrics is controversial in the data mining community and requires attention for each application [18]. The trade-off between speed and accuracy is considered in the selection process as the computational cost varies for the complexity of the time series objects. Another challenge in selecting the distance measures is the compatibility of measures based on the time series representation method.

### 2.3.5 Clustering Algorithm

Clustering algorithms handling time series data are separated into two main approaches: conventional and hybrid. Conventional methods of time series clustering include partitioning, hierarchical, density-based, and model-based algorithms. This subsection will briefly introduce each of the traditional cluster methods and an example of a hybrid method.

#### 2.3.5.1 Partitioning

Partitioning methods of clustering decompose the data into a specified number of partitions,  $k$ , by using an iterative relocation technique. Most partitioning-based methods use a distance metric to group the data elements into clusters. Partition methods specify the number of groups before clustering, which can either be obtained from domain knowledge or through optimization of clusters numbers. Partitioning methods can be associated with high computational costs with large datasets due to the enumeration of partitions. The specification of the partitions limits the data-driven implementation of this method without domain information. Analytical methods avoid the partition limitation by specifying multiple partitions and selecting the best clustering quality result. The two main partitioning clustering algorithms are  $k$ -Means and  $k$ -Medoids.

The  $k$ -means algorithm is a centroid-based method where the mean of the elements represents the cluster. The algorithm groups elements by minimizing the distance between the other elements in the set with respect to the cluster centroid. The  $k$ -Medoids algorithm is a medoid-based method meaning that the prototype element associated with the minimum dissimilarity between the other members represents the cluster. The two partitioning algorithms differ in that  $k$ -Means aims to minimize the cluster sum-of-squares while  $k$ -Medoids try to minimize the sum of distances between each point to the medoid. Furthermore,  $k$ -Medoids selects a cluster member as the medoid while  $k$ -Means selects the mean as the center, which may not be represented in the data. The differences in the cluster representation of the  $k$ -Means and  $k$ -Medoids algorithm is illustrated in Figure 2.13.

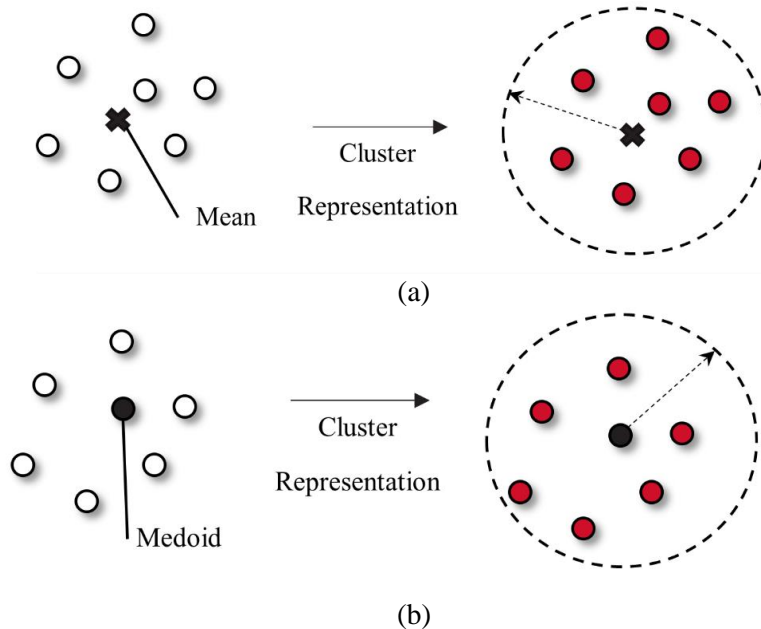


Figure 2.13. Illustration of the clustering representation of (a)  $k$ -Means and (b)  $k$ -Medoids.

The use of a cluster member as the medoid allows any distance metric to be used for  $k$ -Medoids as well as more robustness to noise and outliers [37]. The methods of  $k$ -means and  $k$ -medoids produce "hard" clusters, meaning that each element can only be a member of one cluster.

### 2.3.5.2 Hierarchical

Hierarchical methods decomposed the data set using a hierarchical decomposition resembling tree methods. Hierarchical methods are helpful in applications where additional visualization of the clustering procedure is beneficial through dendrograms, which illustrates the tree structure.

The formation of hierarchical decomposition can either be bottom-up (agglomerative) or top-down (divisive). Divisive methods begin by grouping all elements into a single cluster and iteratively subdividing the members into corresponding clusters. Conversely, agglomerative methods form clusters by initializing each element as an individual cluster and iteratively merging these clusters into larger groups. A similarity measure or distance conducts the merging of clusters, and each cluster is ensured to have at least one member due to the bottom-up nature. Agglomerative methods suffer from scalability limitations for larger datasets due to their quadratic complexity and the restriction of changing clusters once assigned [38]. Agglomerative hierarchical clustering is more widely used because of the increased difficulty of finding optimal splits for divisive clustering. Agglomerative clustering only considers  $n/2$  merges in the first step compared to the  $2^n - 1$  possible splits for top down methods.

### 2.3.5.3 Density-based

Density-based methods produce cluster based on the density of the data subspace, with clusters of dense objects separated by low density subspaces. Density-based methods cluster based on a representation of density in the data alternatively from partition and hierarchical methods, which cluster based on spherical-shaped cluster distribution. Density-Based Clustering Based on Connected Regions with High Density (DBSCAN) assigns clusters based on the density of neighbouring elements [39]. The DBSCAN method has several advantages for the application to flow field data, such as the algorithm does not require prior

information on the number of clusters. Furthermore, the method determines clusters of arbitrary shapes in contrast to the circular cluster shapes of other methods, which benefits the high dynamics of the wake.

#### 2.3.5.4 Model-Based

Model-based clustering algorithms cluster based on recomposed models fitted to the cluster data that represent the entire dataset. Gaussian Mixture Models learn a probabilistic representation of the data by splitting the dataset into numerous Gaussian probability distributions with varying means and covariances [40]. Self-Organizing Maps (SOM) is an unsupervised learning method using neural networks first introduced by Kohonen [41]. SOM models the inherited structures in the dataset by mapping them into a low-dimensional feature space where clusters are extracted.

#### 2.3.5.5 Hybrid Methods

A subset of the clustering analysis research is hybrid clustering which implements multiple clustering techniques to improve results. Most methods implement the costly similarity in shape distance metrics with a reduced dataset known as subclusters represented by a prototype [42], [43].

Aghabozorgi et al. [44] proposed a two-stage hybrid clustering algorithm that produced subclusters based on similarity in time using a hierarchical algorithm. At the data reduction stage, the authors' first step produced a dataset of the subclusters generated by the Cluster Affinity Search Technique (CAST) clustering algorithm to group first-level data. The clusters were generated based on the similarity in time, measured using Euclidean distance. Each subcluster is represented by a time series prototype which is the most typical time series from the subclustered set based on the affinity factor. The next phase, referred to as the clustering step, grouped the subclusters into the final sets. The subclusters were clustered using the  $k$ -Medoids method by computing the distances similarity in shape. The similarity in shape was determined using dynamic time warping on the reduced dataset of prototypes. Based on the subcluster labels of the entire dataset, the final clusters could be worked backward from the final stage. The varying phases and subclusters generated in the analysis are shown in Figure 2.14.

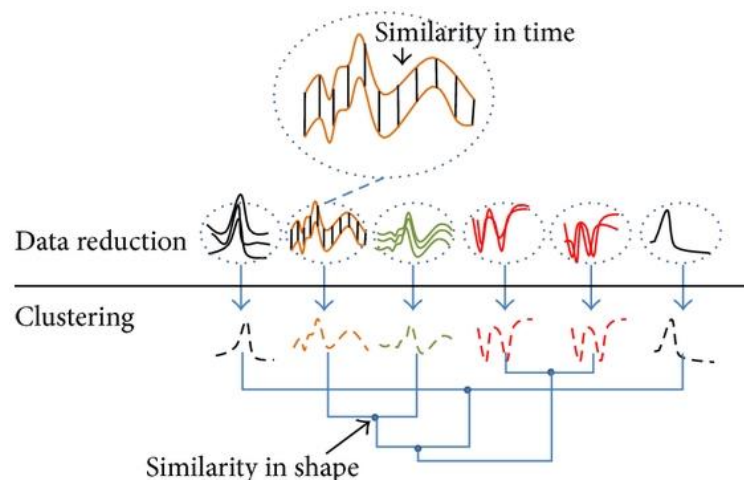


Figure 2.14. Clustering phases implemented in the two-phase method [44].



### 2.3.6 Evaluation Metric

Evaluating clustering techniques based on the generated clusters is vital for quantifiable and meaningful results. The quality of clusters can be evaluated using visualization and scalar accuracy measures.

Scalar accuracy indices evaluate cluster quality by generating a numerical value that represents the accuracy and validity of the classified clusters. Scalar measures comprise external and internal indices. An external index measures the quality of the generated cluster based on the reference to external class value labels of the actual cluster distribution. External indices cannot be used for unsupervised clustering tasks due to the absence of external information leading to the introduction of internal indices. Internal evaluation indices use the intrinsic information of the dataset to determine clustering performance and are referred to as an unsupervised metric since external data are not required.

The clustering quality is typically scored using the internal index based on two parameters: compactness and separation. Compactness measures how close the objects within the same cluster relate to each other, known as intra-cluster similarity [45]. Separation measures how well separated a cluster is from other clusters [45], known as an inter-cluster similarity. Many internal indices are proposed in the literature, such as Calinski–Harabasz, root mean square standard deviation, semi partial R-squared, R-Squared, Davies–Bouldin, Dunn index, Hubert-Levin (C-index), Silhouette, CDbw-Index [46].

#### 2.3.6.1 Inertia

The most common internal evaluation method is inertia or sum of squared error (SSE). The inertia value is the sum of the distances of all the cluster elements to the cluster's centroid. The intuition of the calculation of inertia for a distinct cluster is illustrated in Figure 2.15.

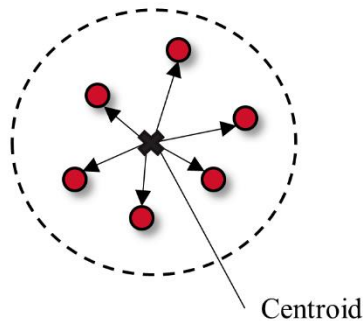


Figure 2.15. Illustration of the inertia metric for a cluster.

The inertia value gives insight into the compactness of the clusters, with a lower bound of zero, indicating that all the elements of the clusters are identical. The inertia of the clustering analysis provides no information on the clusters' separation to other sets, and thus the development of further internal measures has been developed in the literature.

#### 2.3.6.2 Silhouette Index

The silhouette coefficient was first introduced by Rousseeuw in 1987 [47]. The silhouette coefficient measures the similarity of an element to its cluster (compactness) compared to the other sets (separation). Suppose a dataset  $\{D\}$ , of  $n$  objects partitioned into  $k$  groups,  $\{C_1, \dots, C_i, \dots, C_j, \dots, C_k\}$ . The silhouette

coefficient is illustrated with the cluster subsets  $C_i$  and  $C_j$  from the total subspace of clusters  $C_k$  with the corresponding members  $X_i$  and  $X_j$  as shown in Figure 2.16.

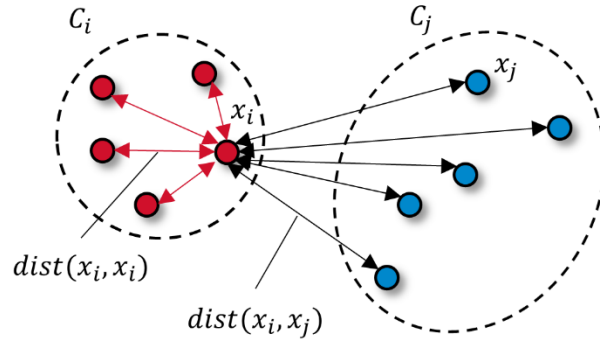


Figure 2.16. Illustration of the silhouette coefficient for two clusters.

The compactness of the cluster is quantified by the parameter,  $a$ , which is the mean distance between a sample and all the points within the same set (compactness). The value  $a$  is calculated for each cluster  $C_k$  and its members  $X_k$  using Equation (2.16).

$$a = \text{mean}_{(C_k)}(\text{dist}(X_k, X_k)) \quad (2.16)$$

The separation of the cluster to the other sets is calculated using the parameter,  $b$ , which is the mean distance from the sample to all the other points in the next nearest cluster (separation). The value  $b$  is calculated for the sample member  $x_i$  to all the members of the other set  $X_j$  using Equation (2.17).

$$b = \text{mean}_{(x_i, C_j)}(\text{dist}(x_i, X_j)) \quad (2.17)$$

The silhouette coefficient is calculated using the parameters  $a$  and  $b$  using Equation (2.18).

$$\text{Silhouette coefficient} = \frac{b - a}{\text{maximum}(a, b)} \quad (2.18)$$

The value of the silhouette index is bounded between -1 and 1. A lower value of  $a$  represents a compact cluster with a lower average distance between the points to a single element in the set. A more considerable value of  $b$  represents clusters with a more significant average distance between a sample element. Therefore, as the silhouette approaches 1, the cluster containing the sample element will be far from the following cluster but compact within the elements of the same cluster. Negative values of the silhouette coefficient represent the sample element being closer to the elements in another cluster than to the assigned elements.

The quality of the clustering analysis can be quantified using the average silhouette coefficient value for all the objects in the cluster, then computing the average silhouette coefficient for the entire cluster dataset.

### 2.3.6.3 Dunn Index

The Dunn index was first introduced by Dunn in 1974 [48]. The Dunn index is the ratio of the minimum distance between clusters (inter) and the maximum distance within a cluster (intra). To quantify the Dunn index, consider the same cluster subsets  $C_i$  and  $C_j$  with the corresponding members  $X_i$  and  $X_j$  as shown in Figure 2.17.

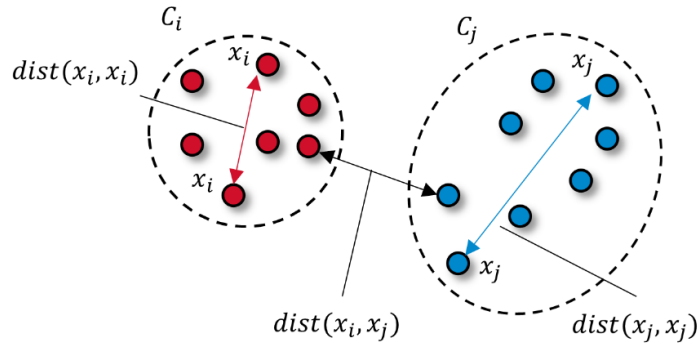


Figure 2.17. Illustration of the Dunn index for two clusters.

The Dunn index is calculated using the distances in Equation (2.19).

$$DunnIndex = \frac{\text{minimum intercluster distance}}{\text{maximum intracluster distance}} = \frac{\min_{(C_i, C_j)}(\text{dist}(X_i, X_j))}{\max_{(C_k)}(\text{dist}(X_k, X_k))} \quad (2.19)$$

The Dunn index should be maximized such that the minimum intercluster distance is more considerable (separated clusters) and the maximum intracluster distance is small (compact clusters).

# Chapter 3

## Literature Review

Since its discovery, vortex shedding of bluff bodies has been an extensive area of research. The research area has recently gained interest with novel wind energy extraction methods from VIV, such as bladeless wind turbines [49] and the Vortex-Induced Vibration Aquatic Clean Energy (VIVACE) generator [50]. This literature review aims to position the contributions of this study, specifically in the generation of vortex shedding maps, and to apply novel data-driven methods for higher Reynolds number cases of vortex shedding.

### 3.1 Vortex Shedding Map Generation

Vortex shedding experiments conducted using forced vibration rely on consistent use of the non-dimensional parameters of the prescribed amplitude and frequency of the oscillations. The vortex shedding map is vital for understanding the cylinder kinematics and its application to energy extraction for VIV generators. The map provides information on energy excitation, which yields the positive energy transfer from the fluid to the body. The positive fluid excitation in the map is critical to ensure the validity of forced vibration experiments application to free vibration cases.

Williamson and Roshko [14] developed the first amplitude and frequency map that studied vortex shedding modes beyond the fundamental synchronization. The authors conducted numerous experiments in a towing tank and identified vortex structures in the cylinder wake vortices using aluminum particles for flow visualization. The experiments sampled the parameter space for the range of Reynolds numbers from 300 to 4000.

Morse and Williamson [7] expanded the investigation by Williamson and Roshko [14] in the parameter space by performing a total of 5680 experimental runs in a water channel that produced roughly 500 hours' worth of data. Digital particle image velocimetry (DPIV) was used to visualize the wake regimes with high resolution (1600 x 1200 pixels) images taken at 10 and 20 milliseconds intervals. The determination of various regimes required a coupled analysis of the flow images and the fluid force measurements by the experimental apparatus. The authors used the abrupt jumps in the fluid forces, which indicated transitions into different regimes, to define the boundaries of the shedding map. For the more complex transition states between vortex shedding modes, the total fluid forces could not identify the modes, which required the investigation of the vortex phase. The methodology of generating the vortex shedding map by Morse and Williamson [7] required large amounts of data in the form of the fluid forces and its decomposition into vortex force and potential force alongside the referencing to the fluid flow images of the wake. The visual inspection relied on an expert to manually identify the spatiotemporal

patterns, which may be possible for low-Reynolds numbers but has limitations for more complex flow regimes associated with higher Reynolds numbers.

The previous studies are the extent of research in generating vortex shedding maps and are limited to low Reynolds number  $Re \leq 4000$ . The lack of investigations in generalizing the vortex shedding behaviour in the parameter space at a higher Reynolds number may be attributed to the more complex dynamics. Wu et al. [16] studied vortex shedding patterns between Reynolds number 35,000 - 130,000 and identified several relatively stable modes, including 2S, 2P, 2PO, P+S, and 2P+4S. Due to instability and variations between cycles, the authors noted the inability to generalize some vortex shedding patterns, including Reynolds number 60,000 - 75,000. Zhang et al. [17] echoed this limitation at the highest free stream turbulence intensity (5%); the vortex structure was indistinguishable in the flow field due to the large amounts of mixing from the increased dissipation energy. The authors found that the turbulence intensity of the incoming stream has a dissipation effect that causes the vortices to become weaker and increasingly difficult to distinguish.

VIV energy harvesting machines are expected to operate at a higher Reynolds number  $Re > 4000$  to achieve feasible energy generation. The benefits of having a vortex shedding map at higher Reynolds numbers are reiterated, but there is a lack of demonstrated methods of generating this map because of the high computational cost and complex dynamics.

## 3.2 Data-Driven Methods for Vortex Shedding

The traditional method of generating vortex shedding maps requires large amounts of data in the form of fluid forces, resolved flow images, and visual inspection from an expert [7], [51]–[54]. Data-driven methods for identifying and clustering distinct flow regimes from high-dimensional, time-resolved flow fields provide a versatile and automated approach to improve confidence intervals while reducing input information.

### 3.2.1 Vortex Shedding Mode Classification

Machine learning models have been applied to vortex mode classification problems using two main methods: identifying global flow structures and classifying local flow signatures. This literature review will focus on the latter due to global methods' limited insight into classifying different vortex modes [55]–[57].

#### 3.2.1.1 Machine Learning using Local Flow Measurements.

The use of local flow signatures for classification is divided into two main applications: namely directly using sensor time-series data and feature extraction from sensor time series into a new feature space. Colvert et al. [58] used local vorticity measurements to train a neural network to classify global vortex structures generated by an oscillating airfoil. The authors performed a multi-label classification task between three vortex modes (2S, 2P+4S, 2P+2S) and achieved minimal error with a network of 34 hidden layers and corresponding classification accuracy of 99%. The work of Colvert et al. was expanded by Alsalman et al. [59], who used the exact local sensor measurements and trained neural network but investigated the effect of different data sensors and sensor array configurations in the flow field.

Wang and Hemati [60] extracted relevant features in the frequency domain obtained from vorticity time series data of a fish-like airfoil shape. The authors computed the frequency spectra from the time series signals using a fast Fourier transform (FFT). The frequency spectrum was then fitted with a

Gaussian bell curve to produce the features used for their wake classification: namely, the mean frequency,  $\mu$ , amplitude,  $A$ , and standard deviation,  $\sigma$ . Wang and Hemati used a relatively simple machine learning model of  $k$  nearest neighbours and achieved a classification accuracy of over 90% for all the experiments. The authors demonstrated that by converting the time series data into a different feature space, the classification problem could be simplified by significantly reducing the computational resources at the cost of a slight accuracy reduction. Using the frequency domain signatures allows better implementation for real-time applications since the entire time series data is not required. The model showed promising results but was tested on a simple classification task involving only the 2S and 2P vortex modes. The addition of more complex modes may reduce the separation in the frequency feature space and require more complex machine learning models to achieve a comparable classification accuracy. Additionally, this study was biologically themed conducted on an airfoil to mimic a fish-like body wake generation and not a cylinder as in the bladeless wind turbine.

### 3.2.1.2 Local Measurement Data Selection.

The standard fluid measurement used for wake classification is the magnitude of vorticity [58], [60]. The traditional thinking is that vorticity, a measurement of the local fluid rotation, will provide a clear signature of the rotating wakes. Alsalman et al. [59] compared the results of several neural networks trained on time series data using the vorticity,  $x$ - and  $y$ -components of flow velocity, and flow speed sensors. The authors found that the  $y$  component of the velocity gave a higher classification accuracy than that of vorticity, even for shallow networks. Alsalman et al. demonstrated that measuring the velocity component transverse to the streamwise flow direction provides the best classification signatures. The authors used the entire time series for classification, and the same gains have not been demonstrated using feature extraction methods on an oscillating cylinder.

### 3.2.2 Vortex Shedding Mode Clustering

Recently in literature, the benefits of the data-driven method of clustering on high dimensional time-resolved flow fields have been gaining attention, specifically for identifying and grouping distinct flow dynamics of vortex shedding from VIV.

Huera-Huarte and Vernet [61] applied fuzzy clustering on digital particle image velocimetry (DPIV) data for the vortex shedding of a flexible cylinder. Proper orthogonal decomposition (POD) was used to reduce the dimensionality of the image data to identify patterns. The method adequately identified the 2S and 2P modes at two elevations along the long flexible cylinder exposed to crossflow of Reynolds numbers 1,200 - 12,000. The work by Huera-Huarte and Vernet [61] provides a pertinent application of clustering on vortex shedding structures for an oscillating cylinder. However, the clustering analysis was only utilized to identify the 2S and 2P modes for the free vibration experiment. The parameter space of amplitude and frequency was not explored since the experiment considered the free vibration of the cylinder. Furthermore, the entire vorticity flow field was required for clustering, despite the use of POD to reduce the data dimensions.

Menon and Mittal [62] studied pitching airfoil wake dynamics using clustering methods to identify and track vortices. The main objective of this study was to analyze the force production on the foil from local vortical regions using the force and moment partitioning method (FMPM). The proper application of this method requires the vortex regions near the airfoil to be accurately isolated. The authors used the clustering analysis to isolate and track the vortices primarily in the leading-edge and trailing edge due to their overall effects on pitching airfoils. The authors used the DBSCAN clustering algorithm for the analysis based on its advantages for flow field clustering. The dataset is comprised of the vorticity fields

from forced vibration CFD simulations of an oscillating airfoil. The pitching frequency and amplitude parameter space were sampled extensively for the varying vortex dynamics. The authors demonstrated the utility of this methodology for the investigation of the dynamic influence of the vortex-induced forces.

Another effort directed for clustering vortex shedding modes was conducted by Calvet et al. [63] to cluster the vortex wakes of bio-inspired propulsors. The authors generated the flow fields behind an oscillating foil using CFD at a Reynolds Number of  $10^6$ . The vorticity flow field in the wake was reduced using an autoencoder and then clustered using the  $k$ -means++ algorithm. The autoencoder consisted of a deep convolutional autoencoder to extract features from the images of the vorticity fields. The authors selected the  $k$ -means++ algorithm after comparing the results of  $k$ -Medoids, hierarchical clustering, and DBSCAN, which all showed no improvement over  $k$ -means++. The combination of feature extraction using the autoencoder and clustering based on the latent space produced an exception method for identifying wake kinematics. The autoencoder was first trained on a simple one degree-of-freedom labelled dataset of known vortex shedding modes to quantify the accuracy of the classification. The vortex wake patterns corresponding to the airfoil application included 2P+4S, 2P+2S, and 2S. Even with a low number of latent space parameters, the clustering achieved 100% accuracy of the modes. The trained autoencoder was then applied to an unlabeled dataset of a two-degree-of-freedom oscillating airfoil with more complex modes. In this case, the wake classifications and the number of clusters are unknown. The quality of the clusters was evaluated both on the silhouette index and visual inspection. The optimal number of clusters for this case was determined to be six through the elbow method, silhouette index, and visual classification of the modes. The authors obtained insights on the vortex dynamics from the clusters successfully. The cluster strategy used by the authors is similar to the one implemented in this study in that the clustering technique was determined for a more straightforward problem of vortex shedding and extrapolated to more complex problems. The success of the authors provides a basis for the methodology used in this study. However, the authors highlight the limitations of the investigation, which greatly depend on the input data used to train the autoencoder and clustering algorithm. The authors draw attention that the results may be corrupted for vortex modes associated with larger amplitudes due to the low density of these cases in the training set. The width of the wakes carried a strong influence in the clustering algorithm, so there is a need to train the algorithm to cluster based on parameters of the overall vortex pattern. Furthermore, the autoencoder was tuned for the performance of the simply one degree-of-freedom case, and despite the promising results, the extrapolation of the autoencoders for more complex vortex wakes is an area of concern.

### 3.3 Summary

In reviewing the literature, it was observed that the traditional method used for generating vortex shedding mode maps requires large amounts of data and intensive supervision. An opportunity was identified to implement a versatile and automated data-driven approach that addresses these limitations. Furthermore, there is a lack of proposed methods of generating the vortex shedding mode map at higher Reynolds numbers due to the increased computational cost and complex dynamics.

Previous studies implementing machine learning techniques have demonstrated improvement over the classical method of identifying wake structures from the resolved flow field. The literature shows promising results using local measurements transformed into the frequency domain as feature vectors in vortex mode classification. However, a lack of studies was identified in applying this method to higher complexity vortex modes generated by an oscillating cylinder. Furthermore, the performance gain using the  $y$ -velocity component has not been demonstrated using the frequency feature space for an oscillating cylinder.

The literature review of studies using unsupervised clustering approaches shows promising results for identifying and dissection of vortex shedding modes. However, the methods require extensive input data to compute the entire flow field, which was then reduced using varying feature extraction methods. The literature identified an opportunity to reduce the input data required by using a local measurement approach to cluster the vortex shedding modes. Studies using clustering all shared the similarity of resolving the vorticity field, but there was a lack of studies investigating the benefits of other flow field measurements. The application of clustering techniques to discover varying vortex shedding modes in an extensively sampled amplitude and frequency parameter space has not been demonstrated. Furthermore, the clustering research for VIV has mainly been regarding pitching airfoils, and the same results have not been investigated for the vortex shedding dynamics of an oscillating cylinder.

This study aims to address the deficiency outlined in the literature by comparing machine learning methods for the classification of complex vortex shedding modes using frequency-domain feature extraction and investigating the effects of data corruption for an oscillating cylinder. Additionally, the absence of a methodology of using clustering techniques for the generation of vortex shedding maps will be addressed in this study.



# Chapter 4

## Methodology

The main objective of this study is discretized into three goals which are addressed in the following three chapters. Chapter 5 provides an effective wake classification strategy by comparing various machine learning models trained using differing simulated vortex shedding local flow measurements and corresponding feature noise analysis. Chapter 6 addressed developing a data-driven method by presenting various unsupervised clustering techniques and demonstrating the method's performance through cluster analysis evaluation metrics and reproduction of the benchmark vortex shedding map. The methods presented in Chapter 6 are then extended in Chapter 7 to provide insights on the underlying dynamical regimes of a higher Reynolds number case by generating corresponding vortex shedding maps.

### 4.1 Datasets

The analysis in this dissertation requires the univariate time series signatures of local flow measurements in the wake of an oscillating cylinder experiencing forced vibration. Two main datasets were used in the following sections:

- Chapter 5: Internal dataset for publication [64].
- Chapter 6 and Chapter 7: Vortex shedding in a turbulent wake obtained from Kaggle [65].

The datasets resolved the turbulent wake behind an oscillating cylinder using numerical 2-dimensional computational fluid dynamic (CFD) simulations in OpenFOAMv2006.

The geometry of a cylinder of constant diameter,  $D$ , subject to an incident flow with a mean velocity,  $U$ , in the streamwise (or,  $x$ ) direction was considered in this study. A slice along the  $x$  and  $y$  plane was used to generate the 2-dimensional CFD simulations based on the reduced computational demands and demonstrated the ability to capture the main features of the response. Though the turbulent wake is a 3-dimensional and multi-scale phenomenon, 2-dimensional simulations can capture the vortex shedding patterns, amplitude, phase, and frequency of VIV. The 2-dimensional domain and boundary conditions used for the numerical study are shown in Figure 4.1.

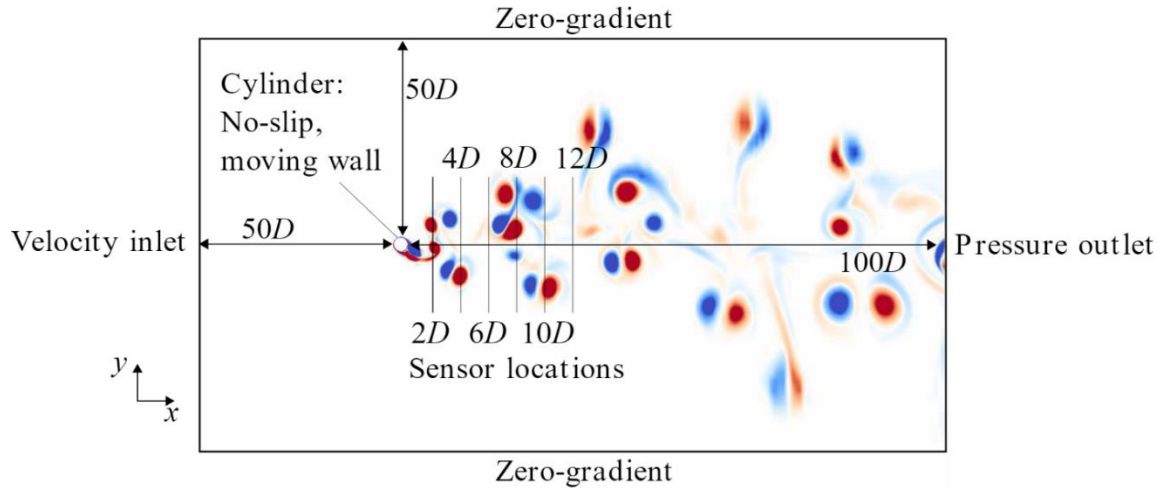


Figure 4.1. Domain and boundary conditions used for the 2D simulation (Not to scale).

The sensors' signal measurements were recorded along six sampling lines located downstream in the wake of the oscillating cylinder, as shown in Figure 4.1. The sampling lines were orthogonal to the  $x$ -axis at streamwise distances of  $2D$ ,  $4D$ ,  $6D$ ,  $8D$ ,  $10D$ ,  $12D$ . Each sampling line recorded flow field measurements at 1000 locations along the line, illustrated in Figure 4.2.

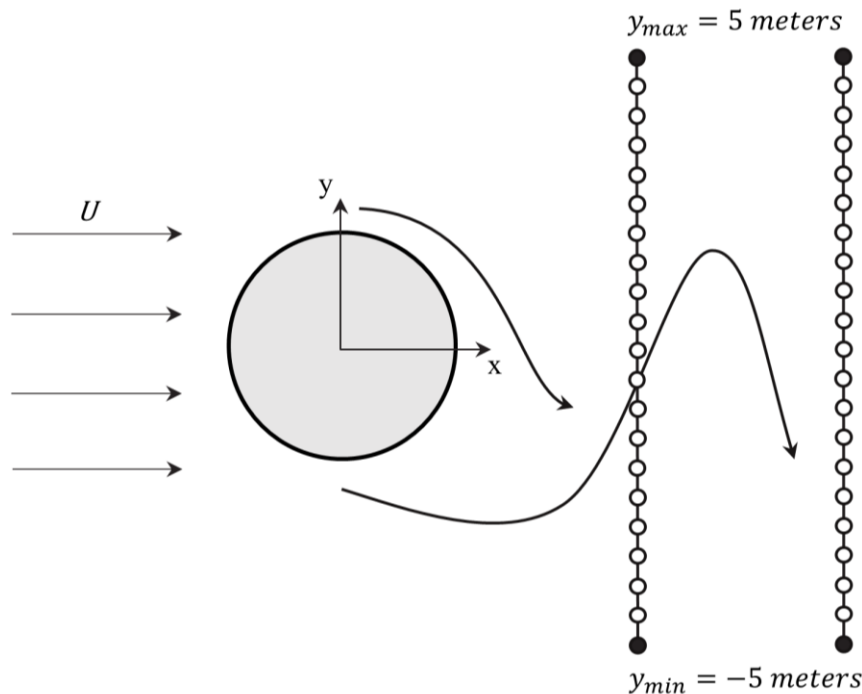


Figure 4.2. CFD sampling lines in the cylinder wake.

From the simulations, the data from four types of sensors were recorded: namely, flow speed,  $|u|$ , the  $x$ -component of velocity,  $u_x$ , the  $y$ -component of velocity,  $u_y$ , and the vorticity,  $\omega$ . The components of velocity in the  $x$ - and  $y$ -directions are derived from the 2-dimensional velocity vector  $\hat{u} = u_x\hat{i} + u_y\hat{j}$ . The flow speed,  $|u|$ , is defined as the magnitude of the flow velocity vector,  $\hat{u}$ . The quantity of vorticity

is defined as the rotation of a fluid element determined by the curl of the velocity vector,  $\hat{\omega} = \nabla \times \hat{u}$ . In this study, the vorticity vector derived from the 2-dimensional flow only has a non-zero component in the  $z$ -direction given by  $\omega = \left( \frac{\partial u_y}{\partial x} - \frac{\partial u_x}{\partial y} \right) \hat{k}$  [66]. An example of the  $x$ -velocity component data signal taken along the sampling line is shown in Figure 4.3 with the corresponding CFD vorticity color map.

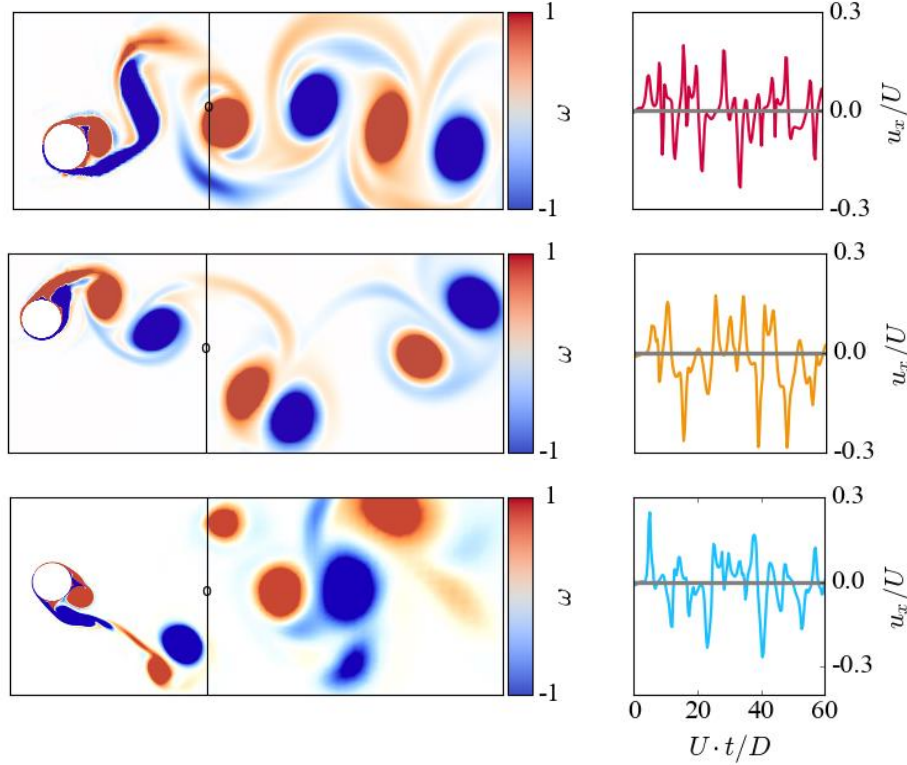


Figure 4.3. Sample vorticity colour maps and  $x$ -velocity component sensor signals for the three wake modes: 2S (top), 2P, (middle), and 2PO (bottom).

#### 4.1.1 Dataset A

The dataset used in Chapter 5 used the RANS equation and  $k - \omega$  shear stress transport (SST) turbulence model. A structured, hexahedral, body-fitted mesh was generated around the cylinder. The total number of cells in the mesh was 147,200. To match Morse and Williamson [7] original experimental setup, the sinusoidal movement of the cylinder is prescribed, and the motion is diffused into the structured mesh. Second-order accuracy was achieved in both spatial and temporal schemes.

The first dataset conducted numerous CFD simulations to generate the flow of three distinct wake modes. The three vortex shedding modes included two wakes modes (2S, 2P) and a transition mode (2PO). The simulations were conducted at the Reynolds number,  $Re = 4000$  (corresponding to  $D = 0.3m$  and  $U = 0.2ms^{-1}$ ) to match the experimental investigation by Morse and Williamson. The non-dimensional groups used to ensure the simulations were conducted for each distinct mode are summarized in Table 4.1.

Table 4.1: Vortex Mode Parameters and Non-dimensional Groups

	$Re$	$A^*$	$\lambda^*$
2S	4000	0.7	4
2P	4000	0.8	8
2PO	4000	0.8	5.8

The sensor measurements were recorded over 500 timesteps with a time step size of 0.5 seconds, resulting in a time series length of 250 seconds. The unit of time was converted into a non-dimensional parameter derived using the free stream velocity,  $U$ , and diameter of the cylinder,  $D$ .

#### 4.1.2 Dataset B

The dataset used for the clustering analysis in Chapter 6 and Chapter 7 used the new schemes of the  $k-\omega$  SST. Numerous CFD simulations were conducted to provide a higher density sampling of the parameter space of forced oscillating outlined by Morse and Williamson [7]. Furthermore, the simulations were conducted at Reynolds numbers 1000, 4000 and 10,000. The Reynolds number data at 4000 was used in Chapter 6 to match the experimental setup of Morse and Williamson [7] and validate the vortex shedding map generation procedure referencing the benchmark map. The Reynolds number data at 10,000 was used in Chapter 7 to extend the map generation method to unknown vortex shedding regimes.

The sensors' measurements varied slightly from the dataset used in Chapter 5 as the time series data were sampled every 0.25 seconds for a total of 100 seconds at the exact locations in the wake. Despite the smaller time series length of the signal, regular patterns are still observed in the data to be used in the cluster analysis.

## 4.2 Mode Classification

The classification of vortex shedding modes utilized the time series local measurements transformed into a reduced feature vector for the subsequent classification using supervised machine learning models. This subsection discusses the preprocessing conducted on the raw time series dataset and the machine learning models used in this study.

### 4.2.1 Preprocessing

The datasets from each sampling line were combined after a generalizability study was conducted to ensure that little information was lost by training a model on the entire set compared to individual sets. The  $k$ -NN model used by Wang and Hemati [60] was implemented,  $k=5$ , based on its heritage in wake classification. When combining the datasets, the testing accuracy only decreased by 8.19% and 0.93% for the  $x$ - and  $y$ - velocity component sensors, respectively. The more considerable loss of the  $x$ -velocity was justified due to the overall lower accuracy results removed in further analysis.

The time series data collected from the CFD simulations were preprocessed to prepare the data for feature extraction and training. The  $x$ -velocity component sensor dataset was preprocessed by removing the mean free stream velocity of  $0.2 \text{ ms}^{-1}$ . Many of the 1000 points along the sample line had minor fluctuations due to the absence of vortex formation. To extract the relevant information from the sampling line, 100 sampling locations were chosen along the line that showed the largest variance from peak to trough in the time series data.

The signals were converted into wake signatures using the protocol developed by Wang and Hemati [60]. The FFT was conducted on the time series to decompose the signal into its frequency domain components. An example of the frequency spectrum of three vortex shedding modes is shown in Figure 4.4.

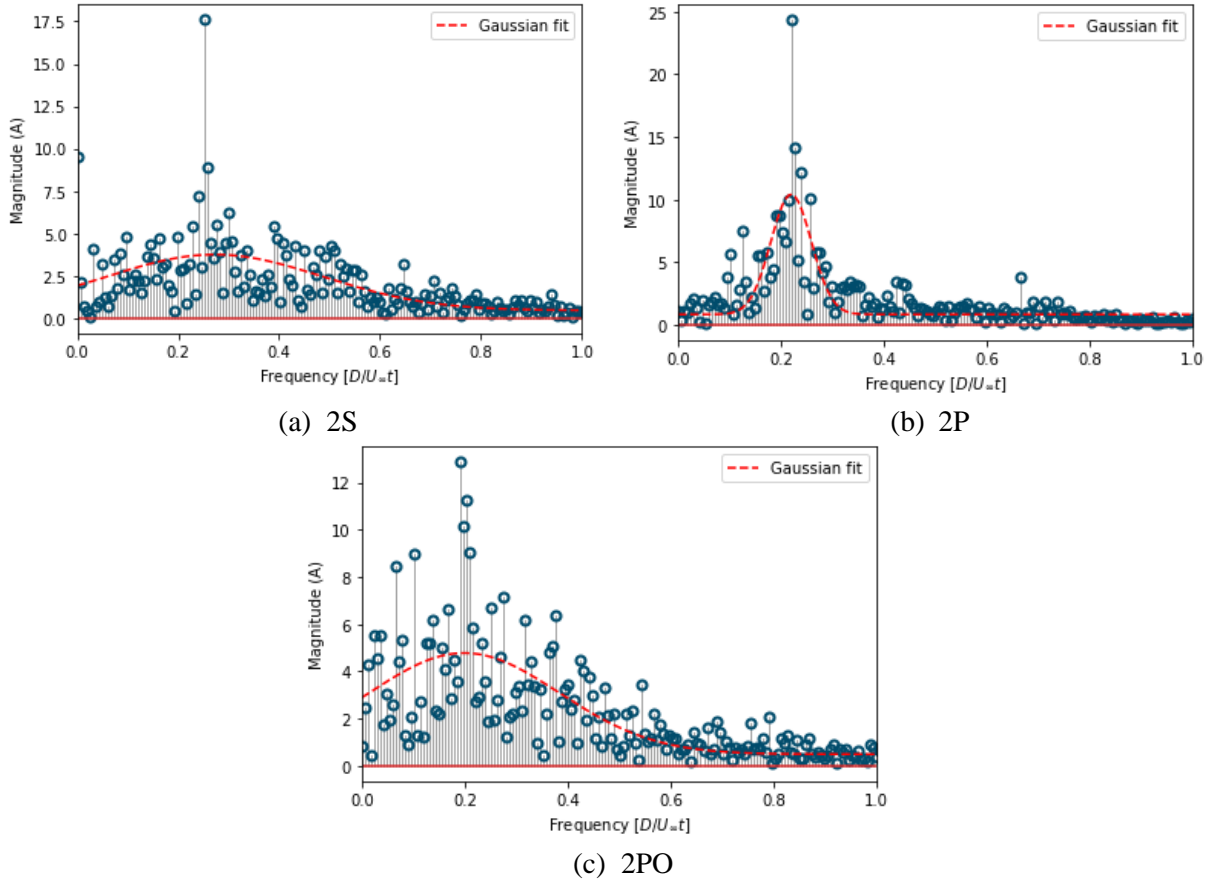


Figure 4.4. Frequency spectrum for (a) 2S, (b) 2P, and (c) 2PO vortex shedding modes.

The feature vector was determined using the Gaussian fit function parameters, which included the mean frequency,  $\mu$ , standard deviation,  $\sigma$ , and the amplitude of the curve,  $A$ . The parameters are shown in a sample case in Figure 4.5.

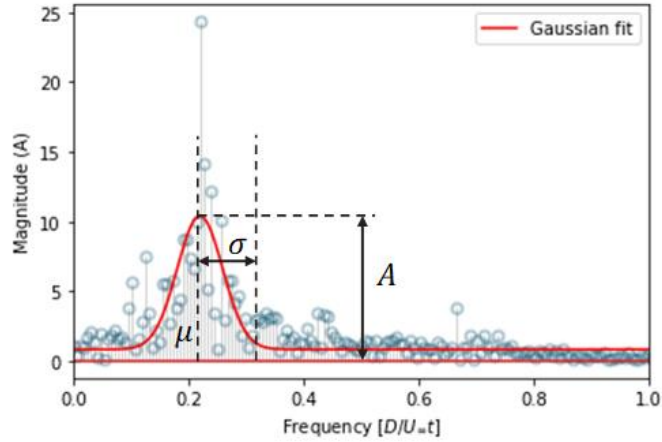


Figure 4.5. Frequency domain feature vector given by the Gaussian fit parameters.

This frequency-domain feature vector combined with vortex shedding mode labels was used as the machine learning models' input.

#### 4.2.2 Machine Learning Models

Machine learning algorithms are founded on using optimization techniques to build a representative model to fit a set of data and, in this case, make label classifications. Supervised machine learning algorithms supply the machine learning algorithms with labelled data from a supervisor/expert. Six supervised machine learning models were investigated for classification: logistic regression, support vector machines, decision tree, random forest, multilayer perceptron (MLP), and  $k$ -nearest neighbours ( $k$ -NN). Traditional machine learning methods were selected based on the objective to provide a robust solution with reduced input data and computational complexity. The six machine learning models offer more straightforward implementations and can achieve solutions with smaller training datasets. Each machine learning model's learning parameters (hyper-parameters) were tuned to determine the best bias/variance trade-off model. Cross-validation was conducted for the hyperparameter tuning stage and repeated on the training set to give insight into the model's generalizability and the predictive performance on the application to new datasets. The investigation was conducted by splitting each dataset into 70/30 training and testing splits. The cross-validation method implemented in this study was 5-fold cross-validation.

### 4.3 Clustering Analysis

The methodology of unsupervised clustering can be subjective, and the quality of the results depends significantly on the specificity of numerous options. The choices selected for the cluster analysis will follow the aspects listed below [38]:

1. Objects to be clustered.
2. Measurements/variables to be used.
3. Standardization of variables.
4. Similarity/dissimilarity measure.
5. Clustering Method.
6. Number of Clusters.
7. Clustering evaluation.

## 8. Interpretation of clusters.

This subsection presents the methodology used to address all of the aspects of cluster analysis to produce meaningful results.

### 4.3.1 Preprocessing

Standardizing the input space is a typical preprocessing step in machine learning applications to enhance performance by transforming the features to the same unit and scale. In time series clustering, standardization can improve the ability of clustering algorithms to produce clusters by removing offsets and amplitude scaling. The standardization implementation depends on the clustering techniques and similarity measures utilized; for example, Gaussian mixture models and the dissimilarity metric Mahalanobis distance are invariant to affine transformations [38].

The implementation of preprocessing techniques should be evaluated for the specific use case and input data since it may introduce biases. For signal data with small variance, standardization and normalization methods may intensify noise with the scaling of the amplitude. In this study, the raw time series signal data is closely centred at zero  $y$  position distance, small relative variance, and amplitude scale are similar between classes. The implementation of preprocessing was determined based on the suitability of the different methods and metrics utilized in this study. Time series data is standardized using z-score (z-normalization), which removes offset translations and amplitude scaling. For the time series  $X_t = \{x_1, x_2, \dots, x_t, \dots, x_T\}$ , z-normalization is defined in Equation (4.1).

$$X_t' = \frac{X_t - \mu_t}{\sigma_t} \quad (4.1)$$

### 4.3.2 Streamwise Analysis

The signals' variance was investigated to determine the effect of Reynolds number, streamwise location, and vortex shedding modes on the signal information. For 2S modes, the variance along the sampling lines remains relatively constant between varying Reynolds numbers. The sensor variance along the sampling lines in the wake of the oscillating cylinder is shown in Figure 4.6.

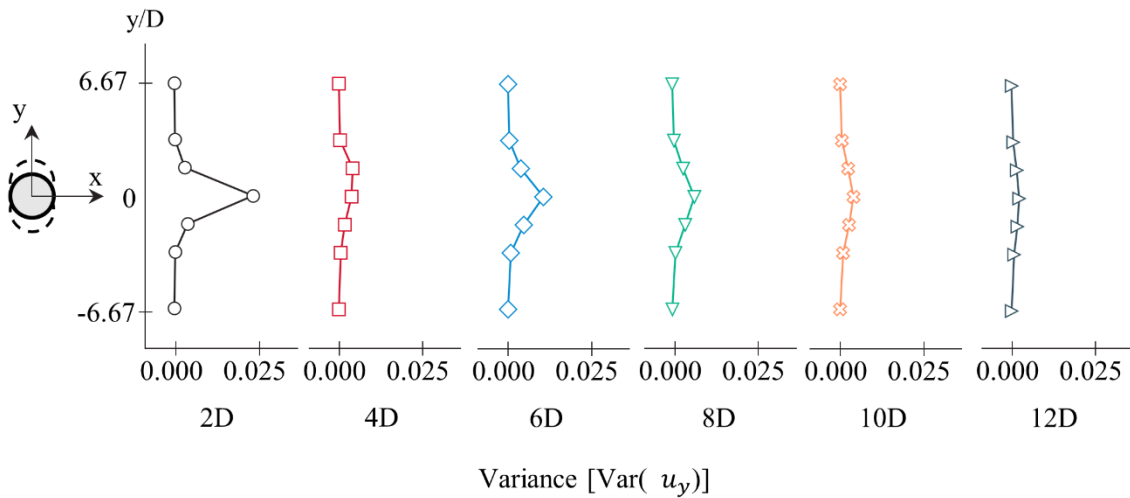


Figure 4.6. Sensor variance of streamwise sampling lines for 2S modes at  $(\lambda^*, A^*) = (4, 0.1)$  for  $Re = 4000, 10,000$ .

The sampling line at 2D shows the most significant spike of variance at the midpoint, where most of the vortex shedding behaviour is present. The sequential sampling of lines further downstream of the cylinder shows a constant dissipation effect. Dissipation results in lower variance, specifically in the 10D and 12D cases where little signal is recorded by the sensors. The relative strength of the signals as the vortex structures travel downstream is seen in the supplementary video provided for the Kaggle dataset [67].

Considering signals that resemble the 2P mode, a more considerable change is observed in the variance behaviour in the sampling lines. The main difference is the more significant dissipation effect of the Reynolds Number of downstream sampling lines for the 2P modes. The variance of the sensors along the sampling lines for the low Reynolds number case,  $Re = 4000$ , is shown in Figure 4.7.

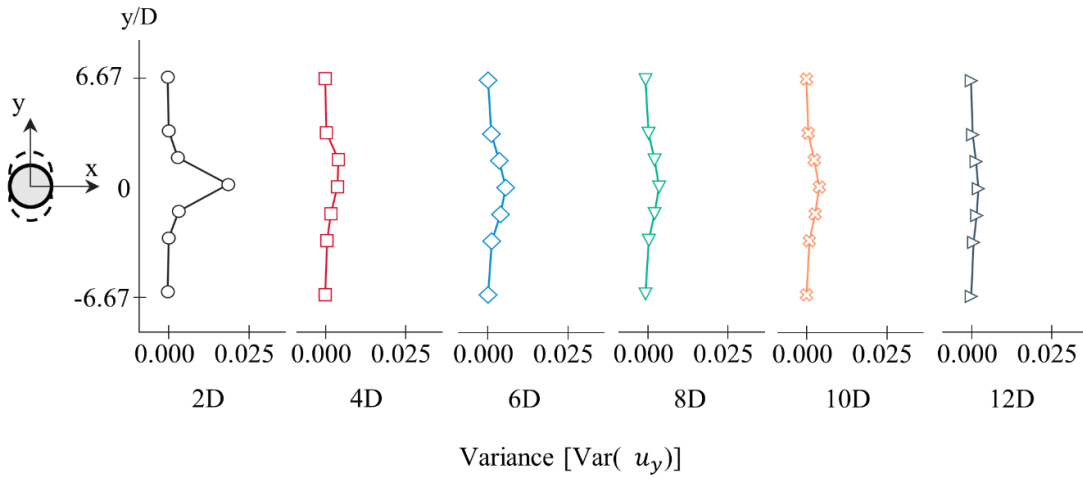


Figure 4.7. Sensor variance of streamwise sampling lines for 2P modes at  $(\lambda^*, A^*) = (6, 0.3)$  for  $Re = 4000$ .

Strong signals are observed for the low Reynolds number case up to 8D and weak signals for the subsequent downstream sampling lines. Conversely, the sensor variance of the 2P mode for the higher Reynolds number case,  $Re = 10,000$ , is shown in Figure 4.8

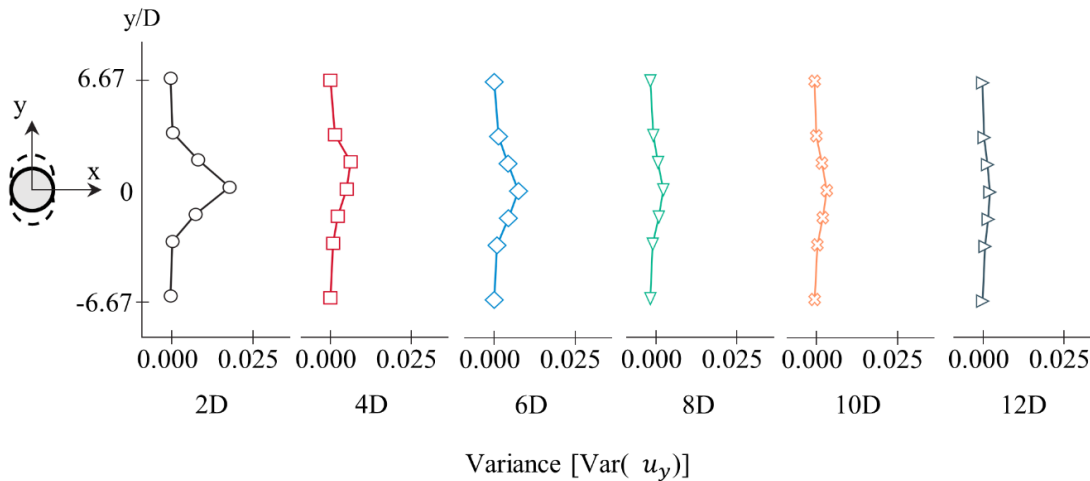


Figure 4.8. Sensor variance of streamwise sampling lines for 2P modes  $(\lambda^*, A^*) = (6, 0.3)$  for  $Re = 10,000$ .



The sampling lines at 2D, 4D, and 6D show relatively strong signals as the vortex structures travel downstream. The subsequent downstream sampling lines show little signal variance starting at 8D from the cylinder.

In summary, for the low Reynolds number case, the data recorded along the sampling line 6D was used for the clustering analysis due to the balance of being far enough downstream to have developed wake mode signals and not too far downstream that the dissipation effect corrupts the signals. For the high Reynolds number case, the data recorded along the sampling line 4D was used for the clustering analysis due to the strong signals of developed wake modes with minimal dissipation effect

### 4.3.3 Subsequence Data Mining

This study's selected time series motif extraction method is conducted using the matrix profile. The matrix profile, first presented by Yeh et al. [68], is a novel algorithm for the time series subsequence all-pairs-similarity-search. The algorithm uses a fast similarity search algorithm under  $z$ -normalized Euclidean distance. The method is simple, parameter-free, and exact, meaning no false positive or false dismissals are provided. The matrix profile has been implemented for various applications since its inception due to its versatility, simplicity, and scalability. These applications include temporary rules of retail product sales [69], ECG anomaly detection [70], and the internet of things for industrial machines [71].

The matrix profile records the distances of a subsequence with sliding window length,  $m$ , to all other subsequences of the same length. The matrix profile contains two components: namely, a distance profile and a profile index. The distance profile is the vector of minimum Euclidean distance, and the profile index is the first nearest neighbour's index. An example of the matrix profile is illustrated with the original time series in Figure 4.9.

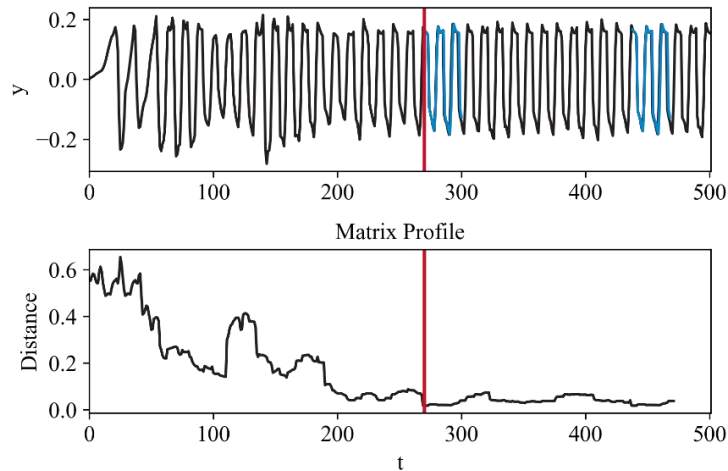


Figure 4.9. Matrix profile example.

The matrix profile can be obtained naively using the computed distance matrix for all pairs of subsequences of length  $m$  of the time series of length  $n$ , as shown in Figure 4.10.

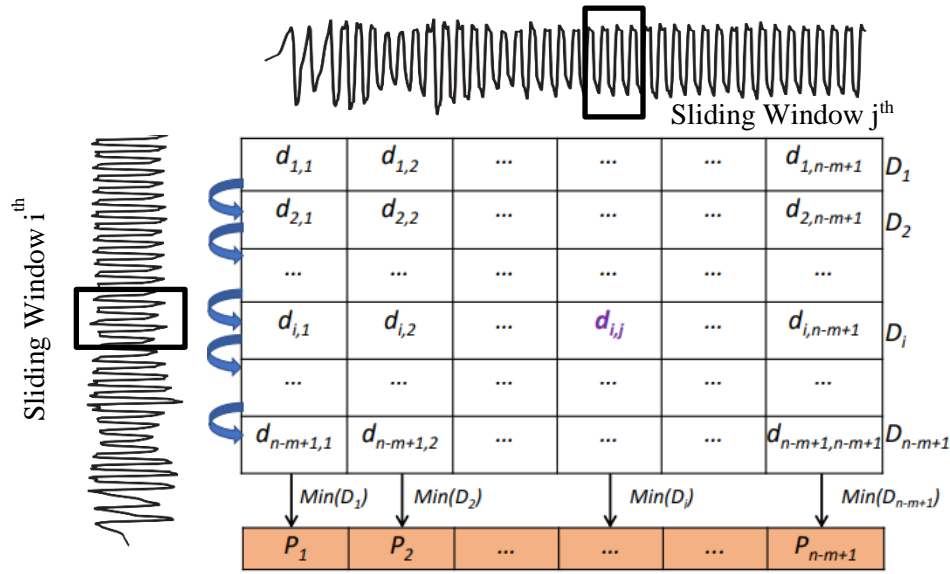


Figure 4.10. Distance matrix to obtain matrix profile, modified from [72].

The minimum distance of each column, not including null diagonal values, produces the matrix profile vector representing the distance between each subsequence and its nearest neighbour. The entire distance matrix would be computationally expensive to compute, and therefore the majority of the work has been developing algorithms for this search [68].

Reading the matrix profile gives an insight into the patterns and anomalies of the time series. Relatively low values indicate that there must be a relatively similar subsequence within the time series. Therefore, low values of the matrix profile correspond to motifs discovered in the sequence. Relatively high values mean that the subsequence has no other subsequence similar in distance and thus is an anomaly. Analyzing the matrix profile for the example in Figure 4.9, the relatively high values at the beginning of the sequence indicate unique subsequences that are not shown in the time series again. The anomaly logically confirms the non-steady-state nature at the beginning of vortex shedding when the shedding mode has not yet been reached. The lowest values in the profile show the prominent subsequence shown in the raw time series and can be extracted as the time series motif for this case. The specified length of the sliding window affects the value of the extracted motif and should be selected carefully for the intended application.

#### 4.3.4 Proposed Clustering Methods

The selection of clustering algorithms is an integral step in the clustering analysis procedure, and special consideration is required for the application and dataset. Clustering algorithms handling time series data are separated into two main approaches: conventional and hybrid. Conventional methods of time series clustering include partitioning, hierarchical, density-based, and model-based algorithms. The algorithms that performed the best for applying clustering time series subsequence of vortex shedding were selected for analysis in this study. The algorithms selected included three traditional single-step and three proposed hybrid methods.

#### 4.3.4.1 Traditional Clustering Methods

Traditional single-step clustering methods were selected based on the demonstrated clustering performance. The traditional clustering methods include the partitioning method of  $k$ -Means, the hierarchical method of agglomerative, and finally,  $k$ -Means applied to the discrete cosine transformed time series data.

The  $k$ -Means method was selected based on its demonstrated ability to extract highly separated clusters. The  $k$ -Means algorithm was implemented using scikit-learn [73] with the  $k$ -means++ initialization method to select initial cluster centers that are distant, resulting in better results than random initialization by avoiding local minimums [74]. The hierarchical method of agglomerative was implemented using scikit-learn [73]. The hyperparameters associated with the best balance of evaluation metrics for the agglomerative algorithm were using complete linkage and cosine affinity distance. The final method of DCT representation using  $k$ -Means algorithm was implemented using the same hyperparameters of the raw time series.

#### 4.3.4.2 Hybrid Clustering Methods

The clustering performance of ordinary methods can be improved using multi-step clustering methods, regarded as hybrid methods [42]–[44], [75]. The objective of hybrid methods is to first conduct a pre clustering phase that produces a large number of separate clusters that are then merged using a final clustering method to provide the number of desired clusters. The hybrid methods perform the best clustering when the silhouette index is maximized in the pre-clustering phase indicating that the high resolution of clusters captures discretely separated clusters. The silhouette index is expected to reduce in the final clustering stage as the clusters are merged to produce more general clusters that provide insight into the overarching patterns, increasing the Dunn index value.

Most hybrid methods use varying distance metrics for each step, usually a similarity in time distance for the pre-clustering and a shape-based similarity distance for the final clustering. The block diagram of the proposed two-step hybrid clustering method is shown in Figure 4.11.

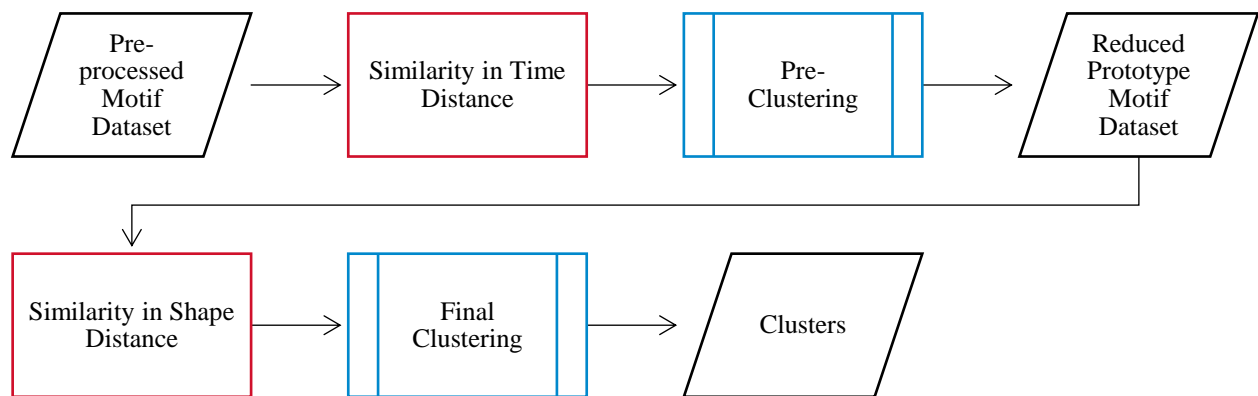


Figure 4.11. Block diagram for the proposed hybrid algorithms.

The advantage of multi-step clustering is the ability to calculate the computationally expensive dynamic time warping similarity matrix on the reduced dataset of prototypes from the first clustering step. The clustering results obtained using the DTW distance provide an advantage for shape-based clustering analysis. An example of the varying clustering stages with the corresponding similarity distances is illustrated in Figure 4.12.

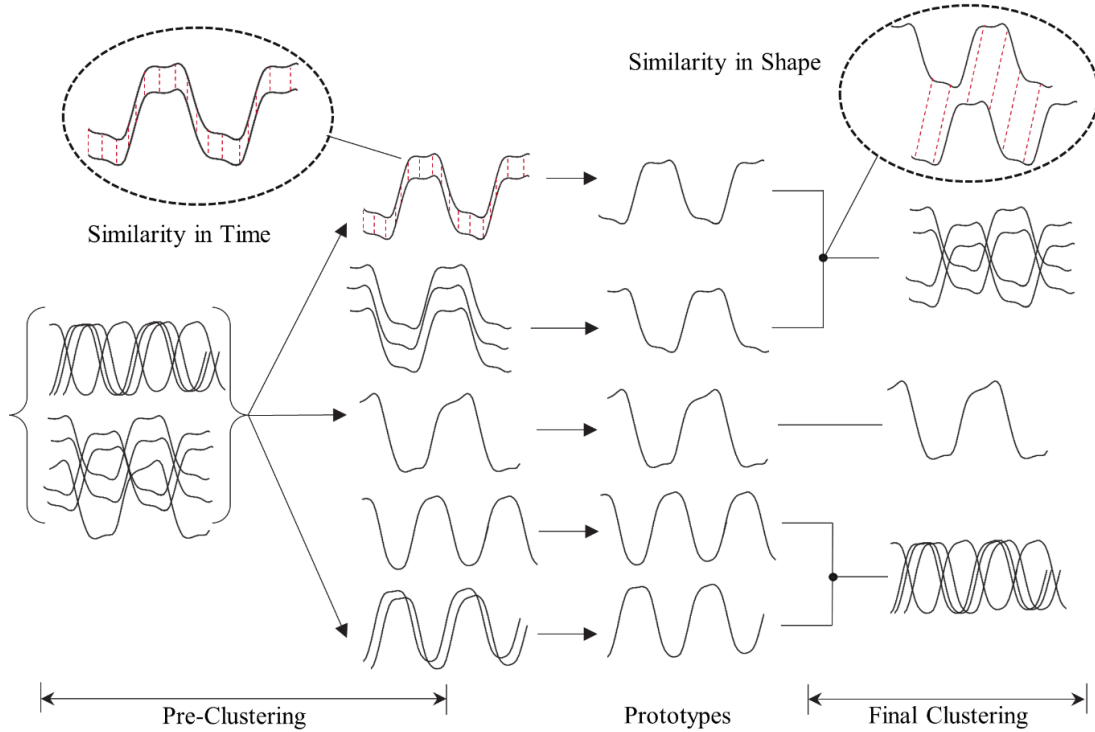


Figure 4.12. Example of the hybrid clustering procedure.

Three hybrid methods are proposed in this study. The first hybrid method, denoted Hybrid Method A, is derived from a similar method proposed by Aghabozorgi et al. [44]. The clustering method implemented by the authors used the Cluster Affinity Search Technique (CAST) clustering algorithm to create subclusters. The subclusters were then grouped further using the  $k$ -Medoids clustering algorithm on the reduced dataset. The CAST algorithm is considered a portioning method based on its sequential clustering approach [76]. The clustering method implemented in this study uses the  $k$ -Medoids clustering algorithm to generate subclusters in the first step based on its advantageous clustering performance.

The following hybrid method, denoted Hybrid B, uses a combination of DBSCAN and agglomerative clustering algorithms. The DBSCAN method provides an advantageous initial clustering step since the clusters are automatically determined based on the input data structures. The DBSCAN method was used to group the entire dataset into subclusters that were then clustered into the final number of groups using the agglomerative algorithm. The hierarchical method of the agglomerative algorithm was selected based on its clustering performance as a single implementation. Other methods performed similarly to agglomerative as a single step, but the hierarchical method had the advantage of less sensitivity to input data. The algorithm was implemented from the clustering module of scikit-learn [73]. The DBSCAN algorithm only requires two parameters: namely, the number of minimum samples and radii. These parameters represent the core samples which is a subset of the data which includes a minimum number of samples,  $min\_samples$ , that are within a distance radii,  $\epsilon$ . The samples that are not

a core sample and are further than  $\varepsilon$  distance from any core sample are considered an outlier by the algorithm. The parameters found using a grid search yielded the best performing algorithm was three radii,  $\varepsilon$ , and five minimum number of samples.

The final hybrid method proposed in this study, denoted Hybrid C, was conceptualized by combining the best-performing single-step clustering analysis into a multi-step method. The hybrid method uses the  $k$ -Means algorithm for the pre-clustering phase and the agglomerative method for the final clustering.

# Chapter 5

## Mode Classification using Machine Learning

This study directly targets the overlapping machine learning and fluid dynamics application of efficiently and accurately classifying wakes identified by Brunton et al. [77]. This section's objective was achieved by quantitatively comparing the various machine learning models trained using differing simulation data and, second, quantifying the effects of data corruption on the models' performance in the feature noise analysis.

### 5.1 Machine Learning Model Performance

The frequency-domain feature space for each of the local measurement sensors is shown in Figure 5.1.

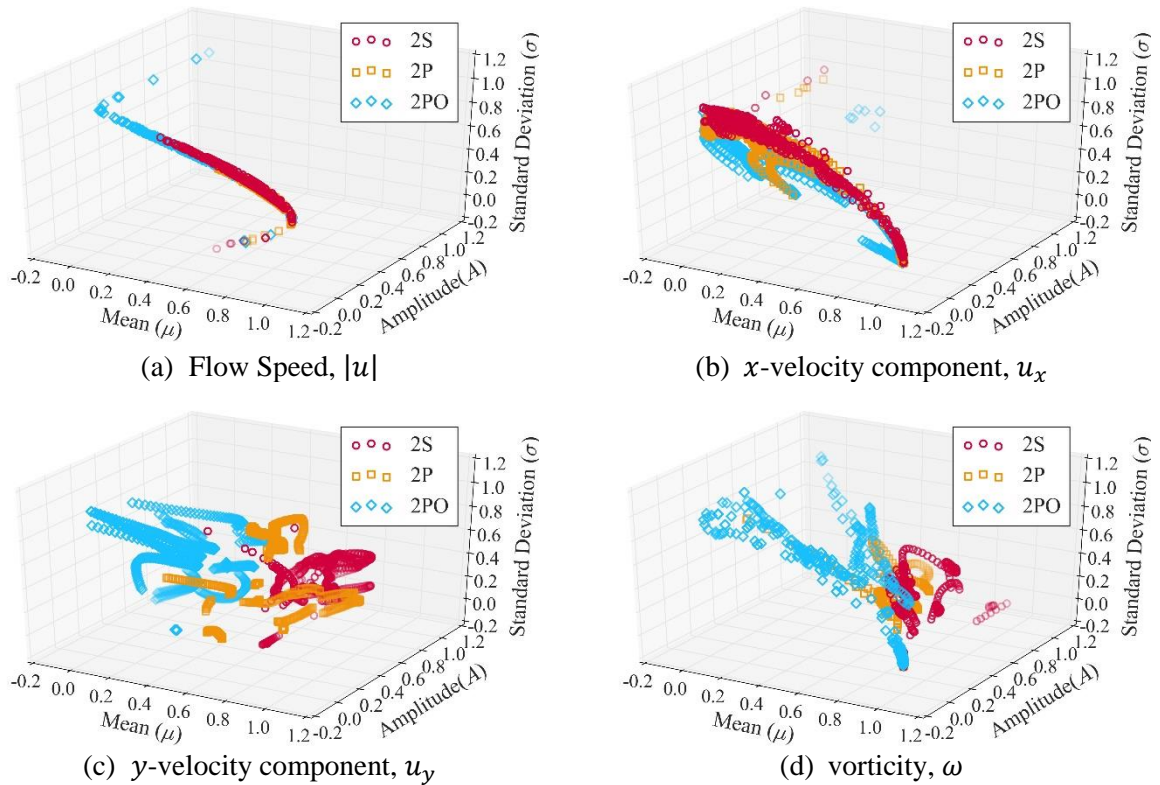


Figure 5.1. Frequency feature space of vortex shedding classes using (a) flow Speed,  $|u|$ , (b)  $x$  velocity component,  $u_x$ , (c)  $y$  velocity component,  $u_y$ , and (d) vorticity,  $\omega$ .

Figure 5.1 shows the  $y$ -component of velocity produces feature vectors that separate the vortex shedding modes the best in the feature space compared to other sensor measurements. The separation in classes aids the models to better identify the signatures of each class which will result in increased testing accuracy.

Several machine learning models were evaluated using the classification accuracy during the cross-validation and testing phases. The cross-validation scores give insight into the models' behaviour with new datasets and reveal the bias/variance behaviours. The mean and standard deviation of the cross-validation scores were plotted for each machine learning model trained using the four local measurements, as shown in Figure 5.2.

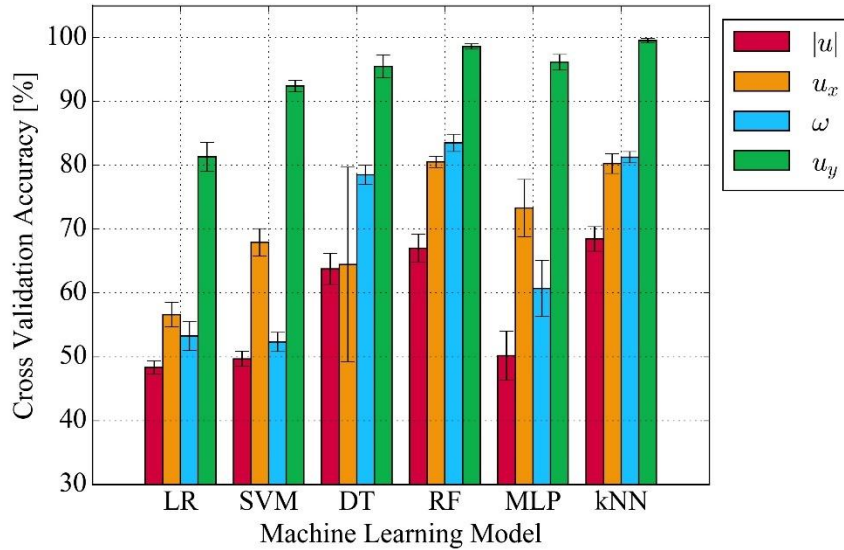


Figure 5.2. Cross-validation scores for each of the local measurement sensor data used to train a variety of machine learning models.

The models that generally performed the best in terms of mean cross-validation scores were the decision tree, random forest, MLP, and  $k$ -NN. Despite the acceptable accuracy scores, the decision tree and MLP models showed the most considerable variation in scores. The larger variation in cross-validation scores demonstrates that these models tend to be more sensitive to input data. Random forest and  $k$ -nearest neighbours performed better concerning the mean accuracy and robustness. Next, the testing accuracy of each of the combinations of machine learning models and input data was evaluated on the testing dataset held out during training. The testing accuracy results are summarized in Table 5.1.

Table 5.1: Testing Accuracy of Machine Learning Models

	Machine Learning Models					
	LR	SVM	DT	RF	MLP	$k$ -NN
$ u $	46.3	48.3	68.3	70.2	56.1	72.4
$u_x$	52.0	69.1	64.4	78.3	76.3	78.0
$u_y$	80.6	91.3	<b>96.7</b>	<b>99.3</b>	<b>98.3</b>	<b>99.8</b>
$\omega$	54.4	53.7	82.2	86.1	75.2	84.4

The  $y$ -component of velocity local measurement sensor dataset demonstrated the highest cross-validation scores and testing accuracies across all the machine learning models, which confirms the results of Alsalman et al. [59]. The improved performance of velocity sensors oriented transverse to the incident flow direction compared to the classically used vorticity measurement may be attributed to the added noise of the  $x$ -component of velocity in the vorticity measurement. In vortex shedding behind a cylinder, the source of the vortices is the generation of the  $y$ -component of velocity since the incident (free-stream) velocity was directed along the streamwise (or  $x$  -) direction.

Overall, the relatively high testing accuracies confirm the viability of using the frequency domain signatures in the application of classifying vortex shedding modes proposed by Wang and Hemati [60]. The input frequency-domain feature vectors resulted in the maximum testing accuracy of 99.3% and 99.8% using random forest and  $k$ -nearest neighbour models. The testing accuracy reported by Alsalman et al. [59] using the entire time series dataset to train a neural network was 100% at a depth of 10 layers. The reduction of 0.7% in testing accuracy obtained in this study was considered acceptable due to the considerable reduction of computational resources and input data required to build the classifier. The  $k$ -nearest neighbours model provided the overall highest testing accuracy of 99.8%, which was an improvement from the 98% accuracy achieved with a similar model by Wang and Hemati. The improved accuracy was attributed to more explicit input data that created a more distinct library of the wake types. The classification algorithms selected to continue analysis were decision tree, random forest, multi-layer perceptron, and  $k$ -nearest neighbours.

## 5.2 Feature Noise Analysis

Machine learning algorithms are built to model the input data used to develop the classifier. The resolution of experimental fluid sensor data will never match that of numerical studies, and increased noise will impact the classification performance for real-world applications. The feature noise analysis evaluates the impact of increasing feature corruption on the selected machine learning models' accuracy. Small transformations were applied to the existing training and testing sets to simulate the data from noisy sensors. The noise was added in varying combinations to the training and testing sets. A control group used for comparison was generated by not introducing noise. Three additional experiments were generated based on varying which sets received noise: namely, Train Clean/Test Clean (CvC), Train Clean/Test Dirty (CvD), Train Dirty/Test Clean (DvC), and Train Dirty/Test Dirty (DvD). A portion of the instances in each dataset was sampled to be transformed. The varying portions were 5%, 10%, 15%, and 20% [78]. Each experiment was conducted five times to demonstrate the variability of the models. The selected models' robustness to data corruption for flow classification is shown in Figure 5.3.



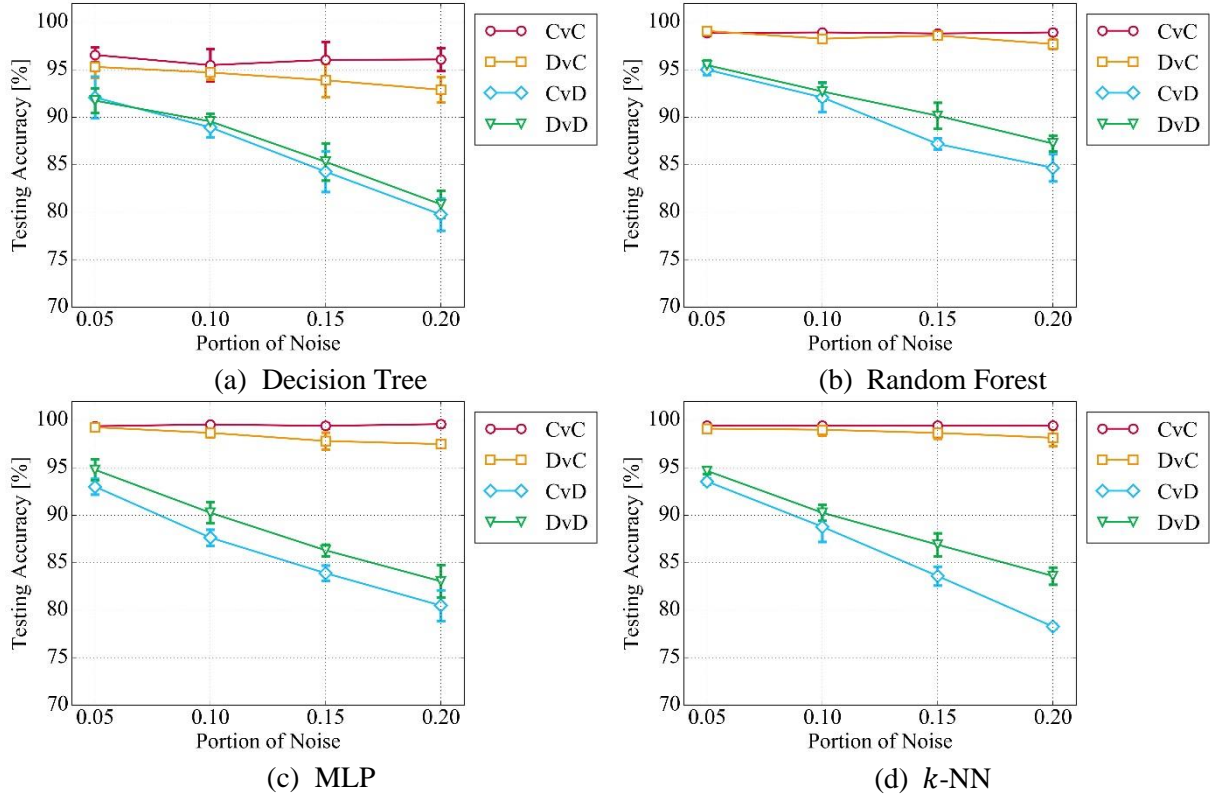


Figure 5.3. Feature noise testing accuracy results for classifiers (a) Decision tree, (b) Random Forest, (c) MLP, and (d)  $k$ -NN

The specific method to include data corruption was achieved by adding a corresponding noise value to each element in the sampled feature vectors. The noise value was randomly selected from a Gaussian distribution fitted to each of the features. Gaussian noise distribution emulates the noisy data from sensors by assuming the error associated with the sensor results in a probabilistic distribution of values centred at the true reading [59].

The testing accuracy results in Figure 5.3 show that an increased level of feature noise corresponds to an accuracy decrease for all the noise cases. The first two cases of CvC and DvC show a relatively constant accuracy for each model over the range of noise levels, with the decision tree classifier most susceptible. The decision tree model showed variance in testing accuracy even with no noise, which is attributed to the algorithm's low bias, high variance nature which tends to overfit input data. The remaining two cases (CvD, DvD) show a constant decrease in testing accuracy with an increased noise level. The importance of testing the models with clean data is apparent with the reduced performance of all the models. The most considerable reduction in accuracy is attributed to the  $k$ -NN model (21.2%) for the CvD case at 20% noise. The reduction of accuracy for the  $k$ -NN model indicates that the  $k$  value selected during clean hyperparameter tuning is too small, making the algorithm sensitive to noise. The  $k$ -NN model after hyperparameter tuning had  $k = 1$  and used the standardized Euclidean distance metric. The random forest testing accuracy reduced the least amount with a maximum reduction of 11.7% for the CvD case at 20% noise. The random forest performs better under attribute noise due to the ensemble algorithm it implements. The algorithm produces an improved model with a low variance while retaining the low bias by aggregating the results of many weaker classifiers with low bias/high variance. After hyperparameter tuning, the algorithm consisted of 107 estimators using the Gini splitting criterion and a maximum depth of 52 for the weak learning trees.

Summarizing the results of the machine model performance and the noise analysis, the random forest classification algorithm was determined to have the best balance of high testing accuracy and robustness to data corruption. Furthermore, the model built using the  $y$ -component velocity provided the best performance in testing accuracy and noise accuracy reduction.

### 5.3 Summary

This section presents an effective wake classification strategy, applying machine learning models trained using fluid sensor data. The results demonstrate the proposed strategy's performance to identify vortex structures from a reduced input feature space accurately. The  $y$ -component of the velocity ( $u_y$ ) achieved the most improved testing accuracy (>15%) compared to the next best quantity, vorticity, which demonstrates the improved feature space separation from the sensor. The highest testing accuracy reported using the  $y$ -component of velocity was 99.3% and 99.8% using the random forest and  $k$ -nearest neighbour models, respectively. The noise analysis on the four best-performing trained models revealed that the random forest algorithm was the most robust to data corruption, with a maximum reduction of 11.7% for the CvD case at the maximum noise level. The importance of higher resolution experimental data for testing the models is apparent from all the models' reduced performance. Combining the results, the random forest classification algorithm (consisting of 107 estimators) was determined as the most advantageous machine learning model due to the balance of testing accuracy and reduced effect from noise.

The method of identifying the wake modes through only the structure of the dataset of local flow measurements is validated through the presentation of this wake classification strategy. The classification task acts as a proof of concept of the validity of the application of the following clustering analysis. Furthermore, the use of the  $y$ -component of the velocity ( $u_y$ ) dataset will provide the best feature separation that is imperative for clustering of time series data.

# Chapter 6

## Vortex Shedding Map Generation at Low Reynolds Number

In this chapter, several unsupervised clustering methods are applied to a low Reynolds number case of vortex shedding from an oscillating cylinder to reproduce the benchmark regime map [7]. Each of the clustering methods is compared based on clustering performance and quality of the vortex shedding maps. The main contribution of this chapter is to validate the clustering methods to reproduce the mode regimes and provide valuable insights on the vortex shedding behaviour at each node.

### 6.1 Methodology

The clustering methods considered include single-step traditional methods and multistep (hybrid) clustering approaches selected based on their demonstrated clustering performance of time series data.

The dataset of extracted subsequences was then clustered using each method and compared using various parameters. First, the generated clusters were validated using visual inspection for the quality and generalizability of the clusters. The clustering performance was quantified using the internal indices of silhouette and Dunn index.

The distribution of clusters was visually displayed using a bi-dimensional embedding of the time series using  $t$ -distributed Stochastic Neighbor Embedding ( $t$ -SNE). The nonlinear dimensionality reduction method of  $t$ -SNE developed by Maaten and Hinton [79] provides an improved map for visualization of high-dimensional data that reveals the data structures at many scales. The datasets are mapped into two components to be displayed in a scatterplot of the various clusters. The two vector components generated by the embedding aims to preserve the structure of the high-dimensional data in the low-dimensional map, specifically preserving the local structures achieved using the non-linear mapping.

Finally, the vortex shedding maps were generated using the proportion of clusters that decomposed each node into the primary and secondary time series signatures identified by the clustering methods. This method of generating the vortex shedding maps provides a detailed description of each mode and the expected time series pattern at each node in the normalized amplitude and wavelength plane.

### 6.1.1 Data Exploration

The dataset obtained from the Kaggle dataset [65] sampled the non-dimensional parameter space along a grid of 5 sampling lines. The following locations on the reference vortex shedding map produced by Morse and Williamson [7] were used to generate the clustering dataset due to observed vortex shedding behaviour in the signals.

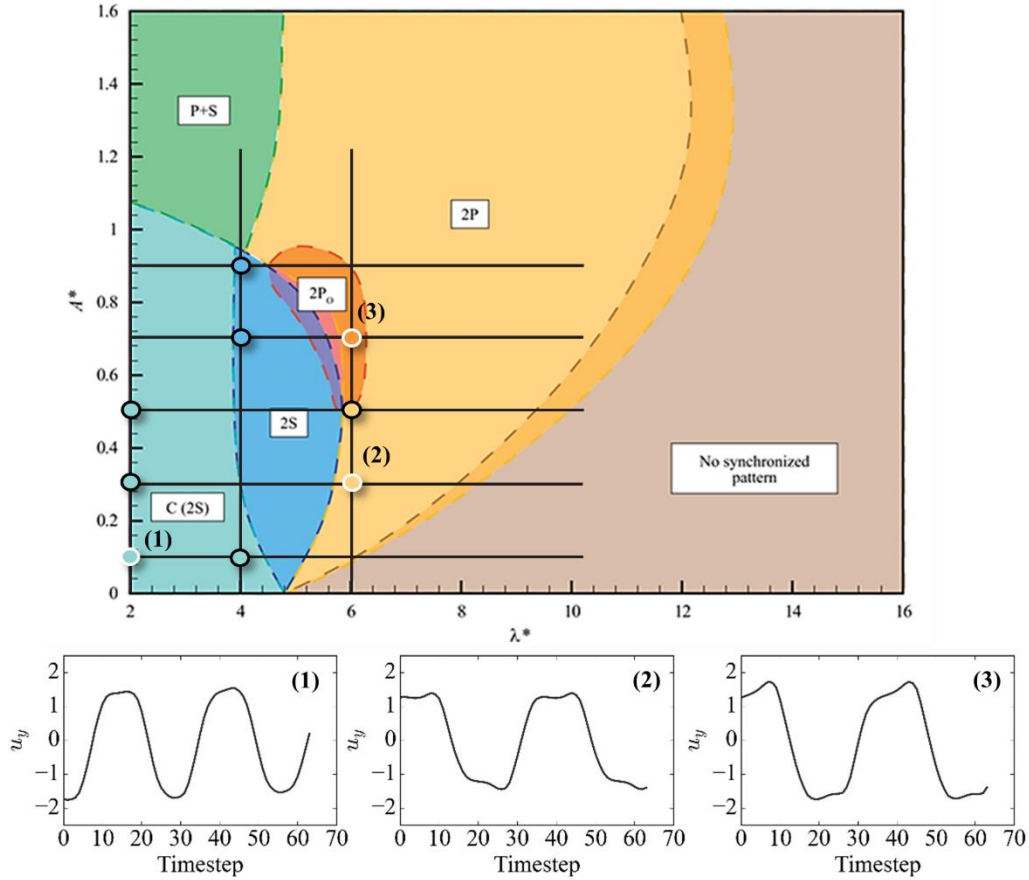


Figure 6.1. Dataset sampled nodes overlaid on reference normalized amplitude–wavelength plane [7].

The sampling nodes were selected for stable vortex shedding behaviour and included three vortex shedding modes: 2S, 2P, and 2PO. Nodes on the frequency sampling line of  $\lambda^* = 4$  was excluded due to the proximity of the points to the C(2S) to 2S transition, where no cohesive patterns are observed in the data. The lack of distinct transition between the C(2S) and pure 2S mode was reported as the least distinct boundary compared to any modes by Morse and Williamson [7]. The data points in the C(2S) region were carefully selected due to the small vortices that coalesce in the near wake detected by the sampling line at 6D.

The expected labels were assigned based on the time series signal's position from the vortex shedding map to produce the expected clusters shown in Figure 6.2. The shaded regions show the maximum and minimum values of the signals over the time steps.

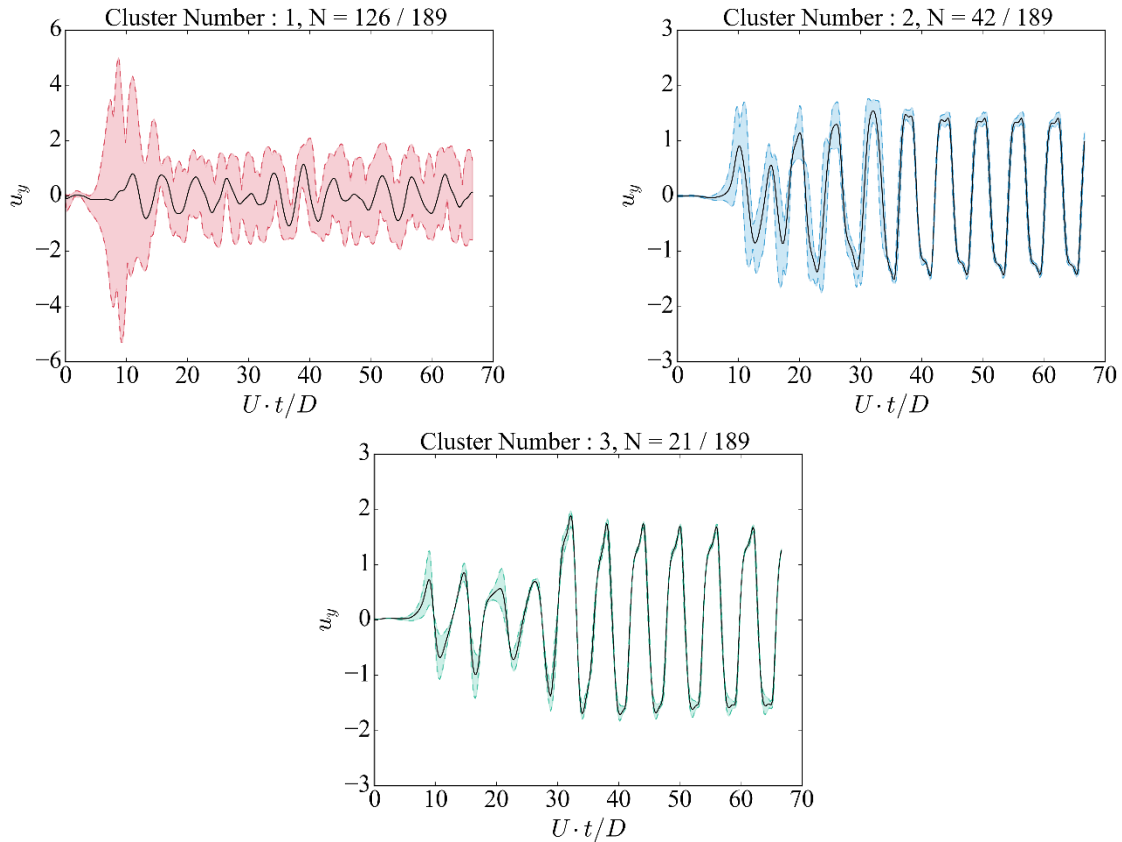


Figure 6.2. Clusters associated with vortex shedding map labels.

Repeated patterns are observed in the time series data shown in Figure 6.2 once the signal reaches a steady state. The subsequence extraction method aims to isolate these repeated patterns for the clustering analysis.

### 6.1.2 Subsequence Extraction

The subsequences in the data were mined using the matrix profile motif extraction method. The specified window size for the algorithm was set to the equivalent of 64 time steps to represent a total of at least two cycles of oscillation. Selecting the sliding window to capture multiple oscillations will improve clustering results by producing more constant and representative patterns. An example of the subsequences extracted using the matrix profile procedure is shown in Figure 6.3.

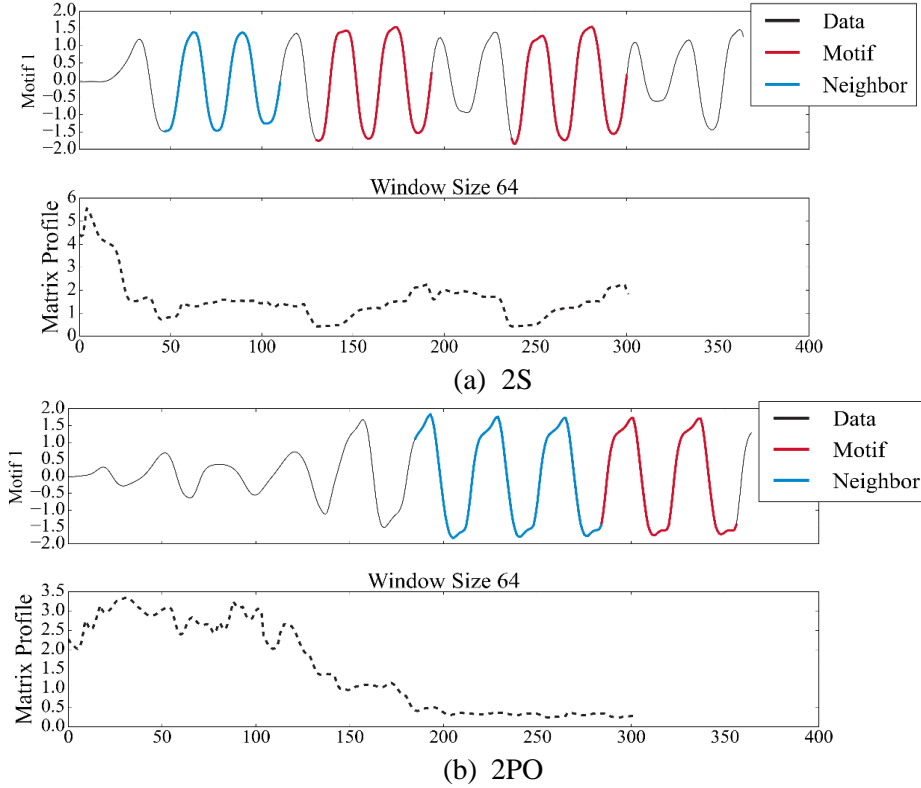


Figure 6.3. Example of motif extraction for signals that resemble (a) 2S and (b) 2PO.

The subsequence extraction is performed on all of the raw time series to extract the single representative pattern for the clustering analysis.

### 6.1.3 Number of Clusters

The number of clusters specified in the cluster algorithms was selected to five, considering the domain knowledge that three distinct vortex shedding modes should be present in the data, as shown in Figure 6.1. The two additional clusters were designated for separating transitional modes such as 2PO and any noise or outlier points identified in the clustering procedure.

## 6.2 Proposed Traditional Clustering Methods

The following section presents the clustering performance of the selected traditional clustering methods through the internal evaluation metrics, cluster plots, latent space cluster distribution, and generated vortex shedding map.

### 6.2.1 *k*-Means

The first clustering method considered was the partitioning algorithm *k*-Means. The clustering performance was quantified using the internal metrics of silhouette and Dunn index, and the results are summarized in Table 6.1.

Table 6.1: Clustering Performance Metrics of  $k$ -Means Method at  $Re = 4000$

Clustering Algorithm	Initialization Method	Evaluation Metric	
		Sil	Dunn
k-Means	k-Means++	0.6559	0.15295

The relatively high evaluation metrics indicate that the clusters generated are sufficiently separate and compact. The quality of the clusters can be visually analyzed from the time series subsequence clusters shown in Figure 6.4.

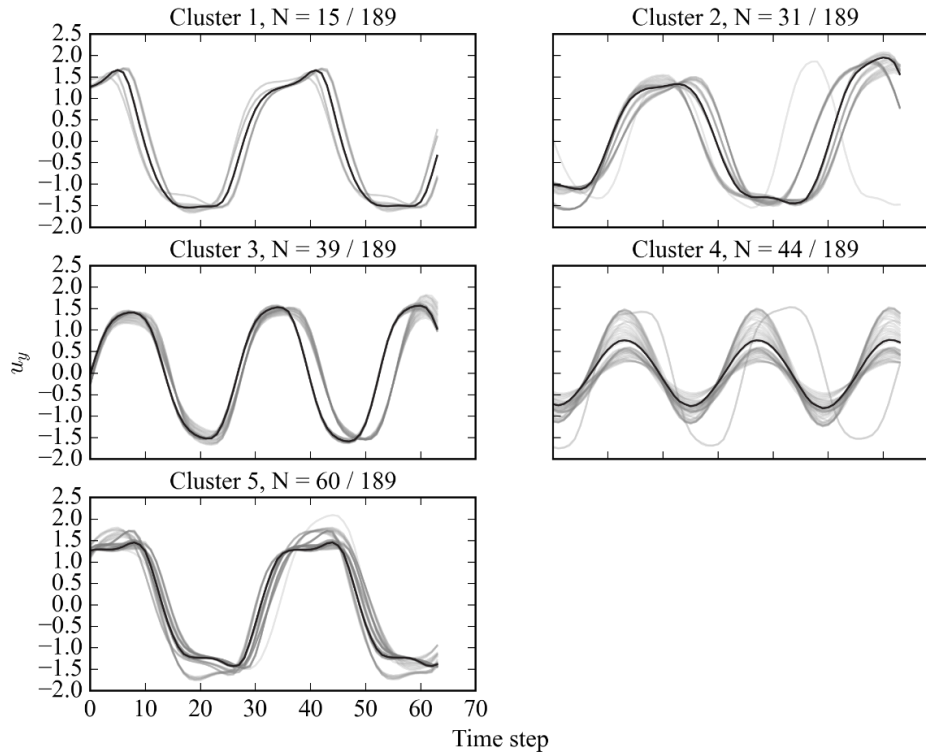


Figure 6.4. Generated clusters by  $k$ -Means method at  $Re = 4000$ .

The generated cluster distribution was visualized using the bi-dimensional embedding method of  $t$ -distributed stochastic neighbour embedding ( $t$ -SNE). The two-dimensional latent space of the generated clusters using  $k$ -Means is shown in Figure 6.5.

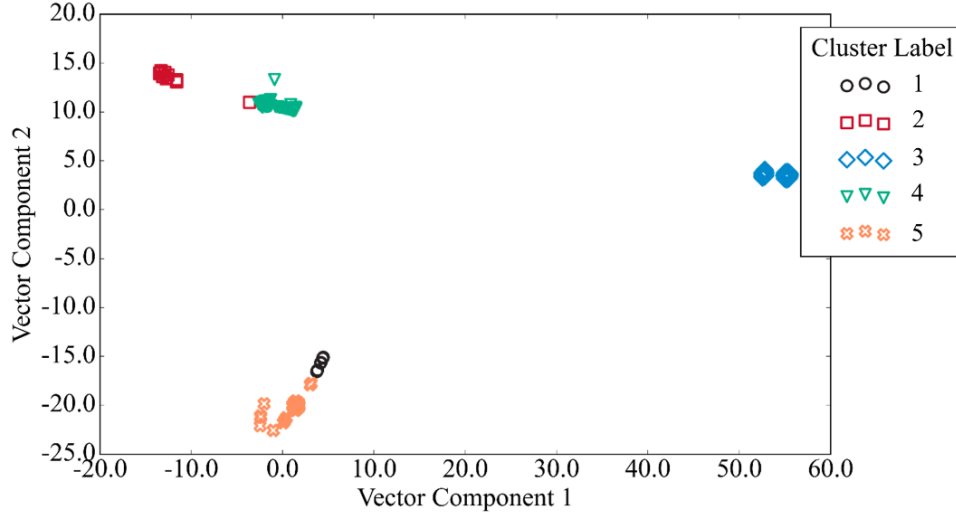


Figure 6.5. Cluster  $t$ -SNE distribution at  $Re = 4000$  using  $k$ -Means.

The similarities of cluster 1 and cluster 5 are shown in the distribution in the  $t$ -SNE space in Figure 6.5. The clusters have a similar double peak behaviour synonymous with 2P and 2PO, but the clustering algorithm isolated the slight ramping of the peak denoted in cluster 1 at the edge of the cluster bunch. The sinusoidal signal denoted in cluster 3 shares little resemblance to the other clusters shown by the sparse subspace the clusters occupy in the  $t$ -SNE space.

The vortex shedding maps were generated based on the proportion of identified clusters candidates at each mode. A primary and secondary vortex shedding mode was identified for each node in the nondimensional amplitude and wavelength space. This method of generating the map allows the identification of strongly clustered nodes and intermittent modes in the parameter space. The primary and secondary clusters identified at each node and the corresponding percentage of each are summarized in Table 6.2.

Table 6.2: Vortex Shedding Map Cluster Candidates for  $k$ -Means at  $Re = 4000$

$\lambda^*$	$A^*$	Cluster Candidate		Candidate Proportion [%]	
		Primary	Secondary	Primary	Secondary
2	0.1	3	4	85.7	9.5
2	0.3	2	5	95.2	4.8
2	0.5	5	2	52.4	47.6
4	0.1	3	-	100	-
4	0.7	4	-	100	-
4	0.9	4	-	100	-
6	0.3	5	-	100	-
6	0.5	5	1	90.5	9.5
6	0.7	1	5	61.9	38.1

The vortex shedding map was then plotted with the primary cluster candidates identified, as shown in Figure 6.6.



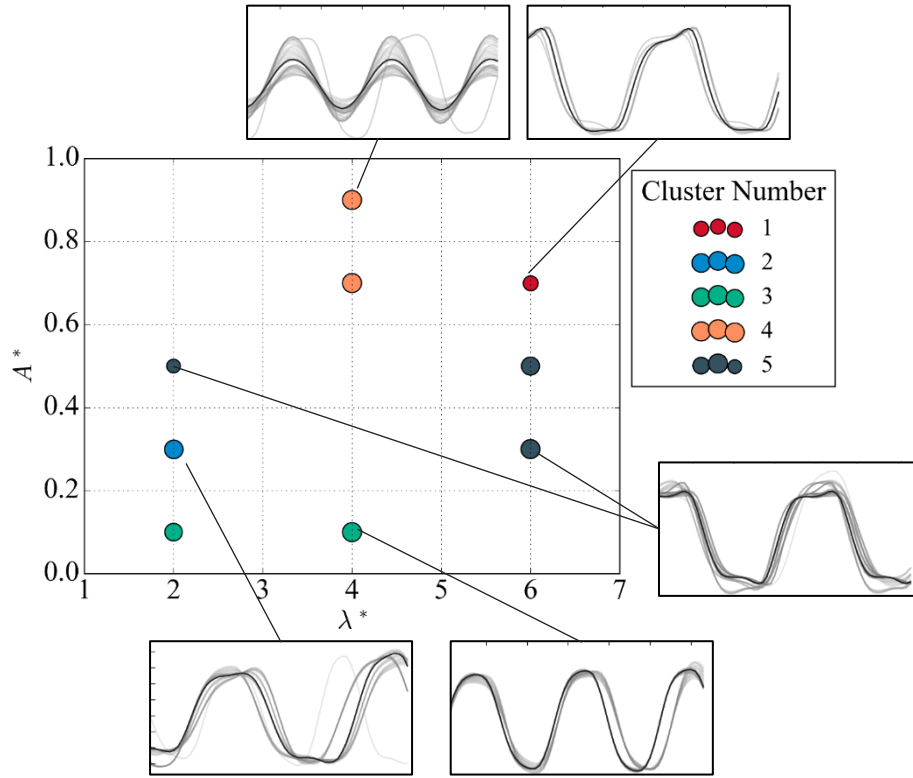


Figure 6.6. Vortex shedding map using  $k$ -Means method at  $Re = 4000$ .

The vortex shedding map produced identified several regions of similar vortex shedding behaviours. First, the generated cluster 1 is designated to the node at  $(\lambda^*, A^*) = (6, 0.7)$  resembling a 2PO mode with a reduced first peak in the pattern. Cluster 2 is primarily for the low-wavelength case  $\lambda^* = 2$  with it occurring mainly at  $A^* = 0.3$  and as the secondary mode at  $A^* = 0.5$ . The low amplitude,  $A^* = 0.1$ , space exhibits the regular sinusoidal pattern indicative of 2S behaviour under the identification of cluster 3. A similar sinusoidal pattern is observed with cluster 4, differing by a lower observed amplitude in the signal. Finally, cluster 5 is primarily located at  $\lambda^* = 6$  at  $A^* = (0.3, 0.5)$  with an additional split located at  $(\lambda^*, A^*) = (2, 0.5)$ . The additional node shows a rather weak identification due to the switching between cluster 5 and cluster 2.

### 6.2.2 Agglomerative

The following traditional clustering algorithm implemented was the hierarchical algorithm of agglomerative. The internal indices used to quantify the clustering performance are summarized in Table 6.3.

Table 6.3: Clustering Performance Metrics of Agglomerative Method at  $Re = 4000$

Clustering Algorithm	Linkage	Affinity Distance	Evaluation Metric	
			Sil	Dunn
Agglomerative	Complete	Cosine	0.6794	0.61721

The agglomerative method produces clusters with a much larger Dunn index indicating the quality of the groups. The clusters associated with the evaluation metrics are shown in Figure 6.7.

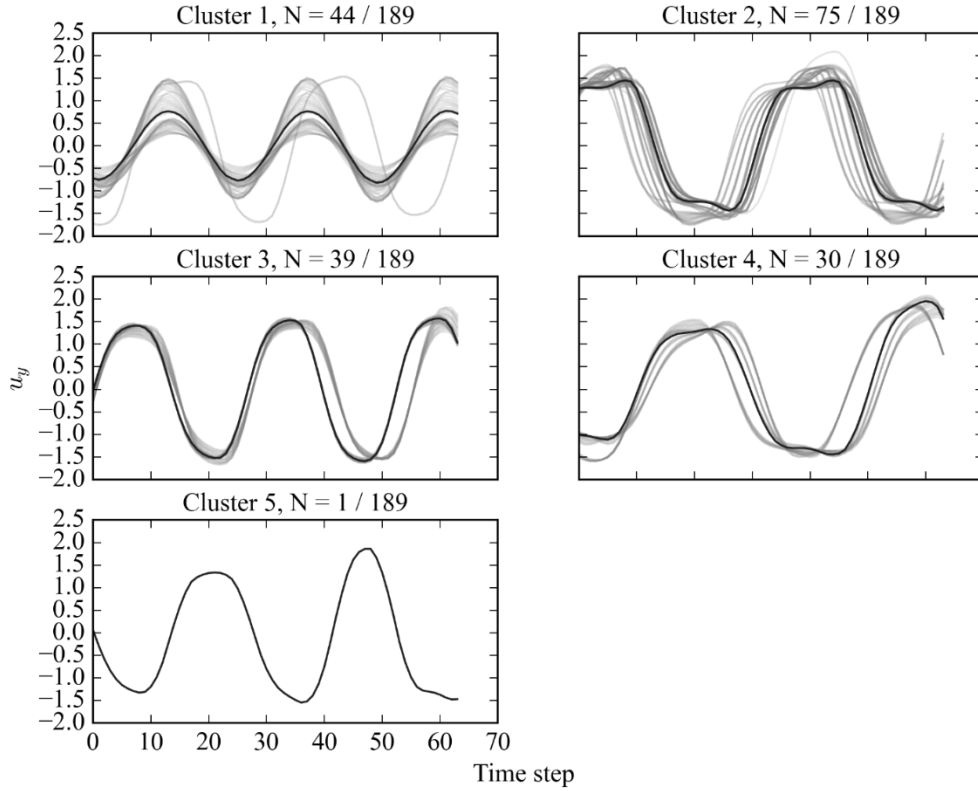


Figure 6.7. Generated clusters by Agglomerative (complete, cosine) method at  $Re = 4000$ .

The hierarchical method of the agglomerative algorithm provides a great visualization of the time series clustering procedure, as shown in Figure 6.8.

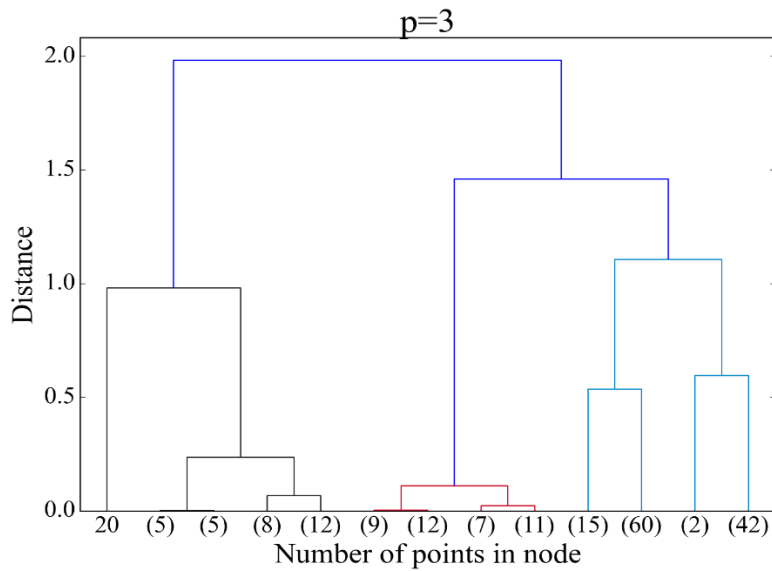


Figure 6.8. Agglomerative dendrogram at level  $p = 3$ .

The algorithm's clustering approach can be seen further in the distribution of clusters in the two-dimensional latent space shown in Figure 6.9.

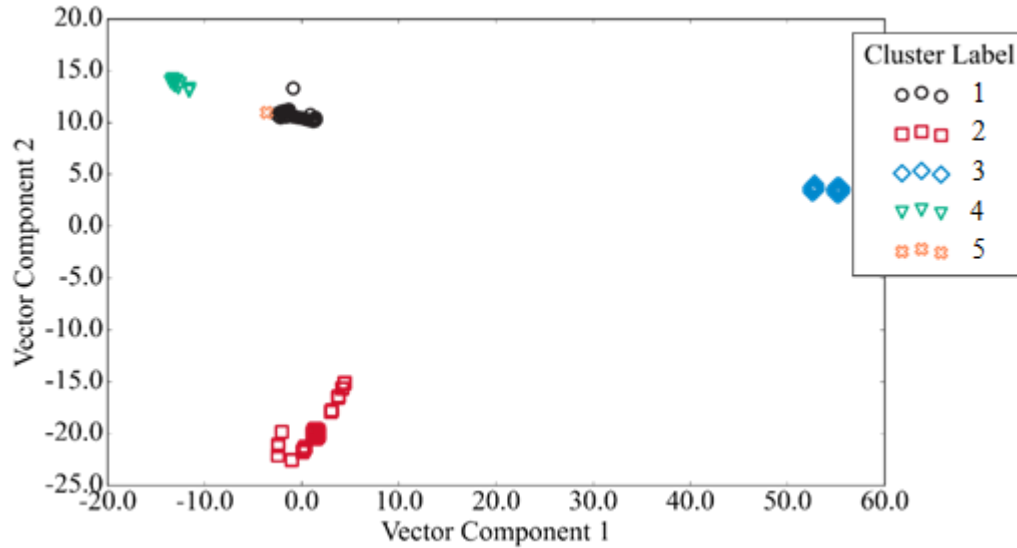


Figure 6.9. Cluster  $t$ -SNE distribution using the agglomerative method at  $Re = 4000$ .

The latent space clearly shows the distinction of the generated clusters, specifically for clusters 2, 3 and 4. The sample selected for cluster 5 inhabits the subspace of cluster 1, which may be due to the higher amplitude and longer wavelength of this sample. Although the sample point denoted as cluster 5 seems to better fit within the cluster 1 subspace, the effect of this class can be observed in the vortex shedding map. The primary and secondary clusters identified at each node and the corresponding percentage of each are summarized in Table 6.4.

Table 6.4: Vortex Shedding Map Cluster Candidates for Agglomerative  $Re = 4000$ .

$\lambda^*$	$A^*$	Cluster Candidate		Candidate Proportion [%]	
		Primary	Secondary	Primary	Secondary
2	0.1	3	1	85.7	9.5
2	0.3	4	2	95.2	4.8
2	0.5	2	4	52.4	47.6
4	0.1	3		100	
4	0.7	1		100	
4	0.9	1		100	
6	0.3	2		100	
6	0.5	2		100	
6	0.7	2		100	

The vortex shedding map was then plotted with the primary cluster candidates identified, as shown in Figure 6.14.

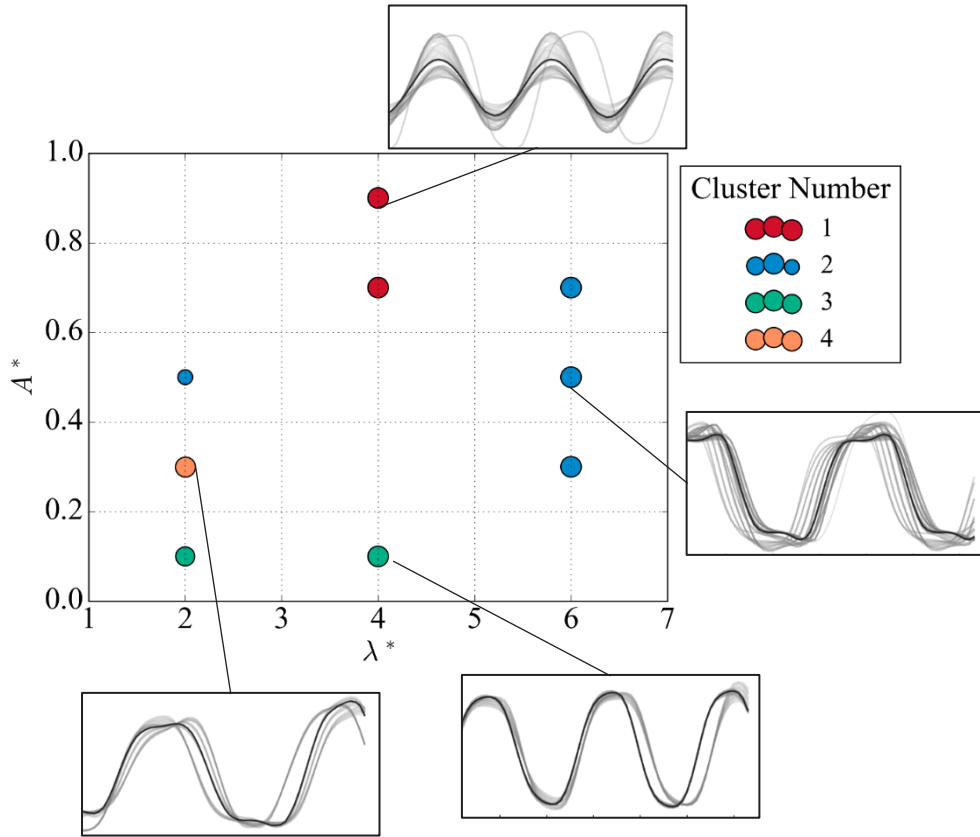


Figure 6.10. Vortex shedding map using the agglomerative method at  $Re = 4000$ .

The vortex shedding map isolates several similar regions of subsequence patterns, as seen in the  $k$ -Means method. The patterns similar to the signature expected for the 2S mode were observed for clusters 1 and 3, only varying by amplitude. A group of nodes identified as cluster 2 on the sampling line  $\lambda^* = 6$  demonstrated the double peak of the pair of vortices being shed from the 2P mode.

### 6.2.3 Discrete Cosine Transform Representation with $k$ -Means

The final clustering method repeated the use of the  $k$ -Means algorithm but was trained on the time series dataset represented using the discrete cosine transform (DCT). The clustering performance for the reduced dataset using the discrete cosine transform is summarised in Table 6.5.

Table 6.5: Clustering Performance Metrics of DCT dataset using  $k$ -Means Method at  $Re = 4000$

Representation Method	Clustering Algorithm	Evaluation Metric	
		Sil	Dunn
DCT	$k$ -Means	0.6559	0.15295

The clusters identified using the  $k$ -Means algorithm trained on the transformed dataset are shown in Figure 6.11.

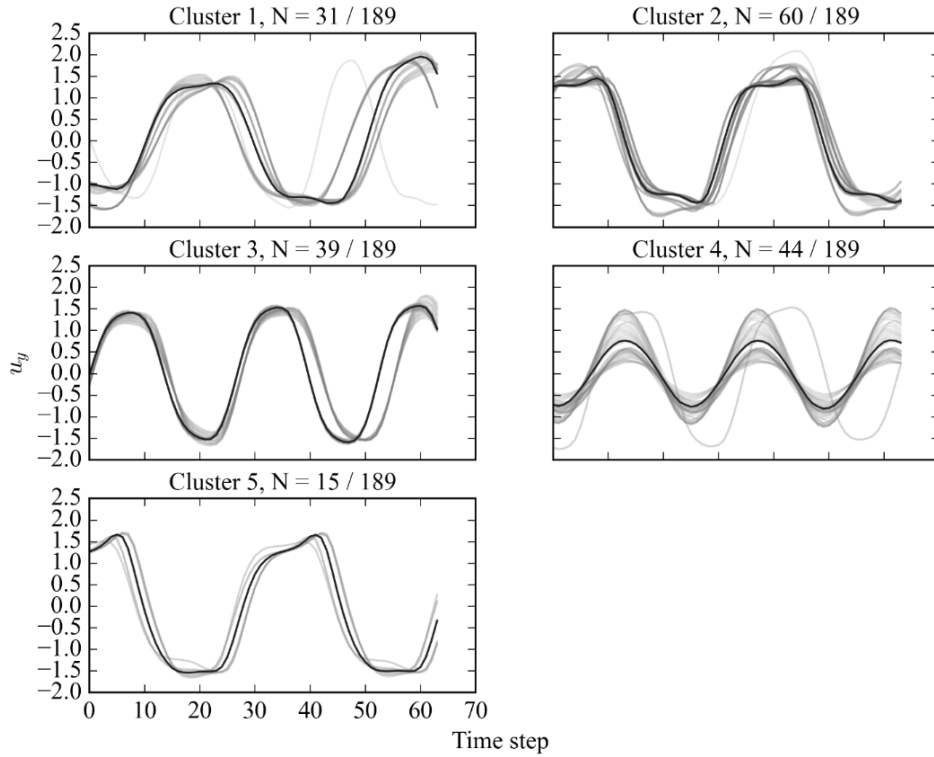


Figure 6.11. Generated clusters by  $k$ -Means method on DCT dataset at  $Re = 4000$ .

The samples of each cluster resemble the results of the  $k$ -Means method on the raw subsequence dataset. Both methods adequately identified the periodic sinusoidal signals of the 2S mode and even the double-peaked signals of 2P and 2PO. The cluster distribution produced in the two-dimensional latent space generated using  $t$ -SNE is shown in Figure 6.12.

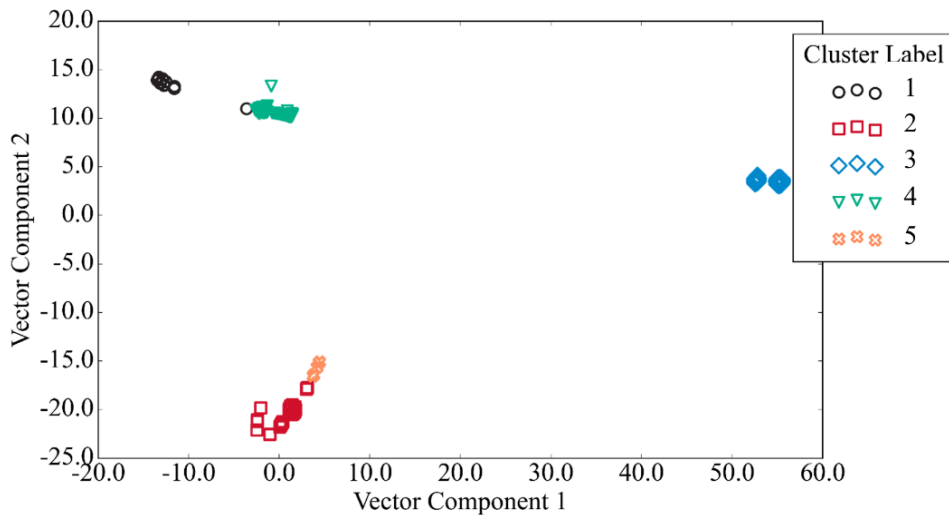


Figure 6.12. Cluster  $t$ -SNE distribution at  $Re = 4000$  using  $k$ -Means on DCT dataset.

The primary and secondary clusters identified at each node and the corresponding percentage of each are summarized in Table 6.6.

Table 6.6: Vortex Shedding Map Cluster Candidates for  $k$ -Means on DCT dataset  $Re = 4000$

$\lambda^*$	$A^*$	Cluster Candidate		Candidate Proportion [%]	
		Primary	Secondary	Primary	Secondary
2	0.1	3	4	85.7	9.5
2	0.3	1	2	95.2	4.8
2	0.5	2	1	52.4	47.6
4	0.1	3		100	
4	0.7	4		100	
4	0.9	4		100	
6	0.3	2		100	
6	0.5	2	5	90.5	9.5
6	0.7	5	2	61.9	38.1

The primary cluster candidates with the corresponding proportion were plotted on the normalized amplitude wavelength plot shown in Figure 6.13.

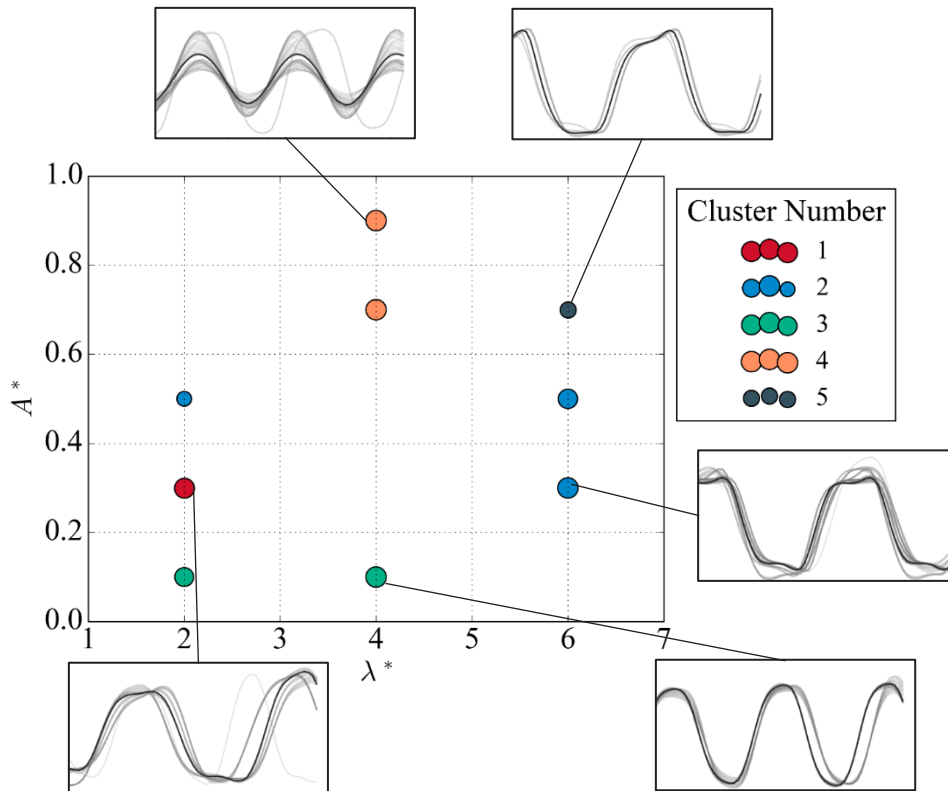


Figure 6.13. Vortex shedding map using  $k$ -Means method on DCT dataset at  $Re = 4000$ .

The vortex shedding map with the DCT representation of the dataset and  $k$ -Means algorithm share many similarities, including the identification of the isolated 2PO mode located at  $(\lambda^*, A^*) = (6, 0.7)$ .

### 6.3 Proposed Hybrid Clustering Methods

Three hybrid methods are proposed in this study for the cluster analysis, and each is compared based on internal evaluation metrics, cluster plots, latent space cluster distribution, and generated vortex shedding map.

#### 6.3.1 Hybrid Method A

The results of the hybrid method A are presented in sequential order of the pre-clustering and final clustering phases. Unlike the implementation of  $k$ -Medoids in single-step clustering analysis, the number of clusters generated in the first step is unbounded, and the optimum number of clusters must be selected. The evaluation metrics of silhouette and Dunn index for an increasing number of clusters are shown in Figure 6.14.

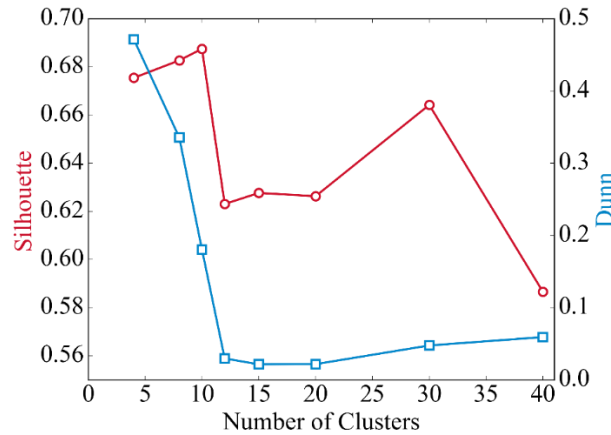


Figure 6.14. Evaluation metrics for the number of clusters generated using  $k$ -Medoids.

The optimum number of clusters was determined to be 30 since the multi-step approach maximizes the separation of the clusters for the first stage.

The final clustering step combines the reduced dataset using the dynamic time warping (DTW) distance and the  $k$ -Medoids algorithm. The number of clusters for the final step is defined as the same for the single-step clustering methods at five. The clustering performance results of both phases are summarized in Table 6.7.

Table 6.7: Clustering Performance Metrics of Hybrid A Method at  $Re = 4000$

Phase	Clustering Algorithm	Number of Clusters	Evaluation Metric	
			Sil	Dunn
1: Pre-Clustering	$k$ -Medoids	30	0.6565	0.05469
2: Final Clustering	$k$ -Medoids	5	0.4971	0.01430

The pre-clustering phase isolates discretely separated clusters identified by the relatively high silhouette index. However, the performance of the evaluation metric decreases for the merged clusters in the final clustering phase. The explanation of the reduced clustering performance can be determined by the plotted cluster samples of Hybrid A, shown in Figure 6.15.

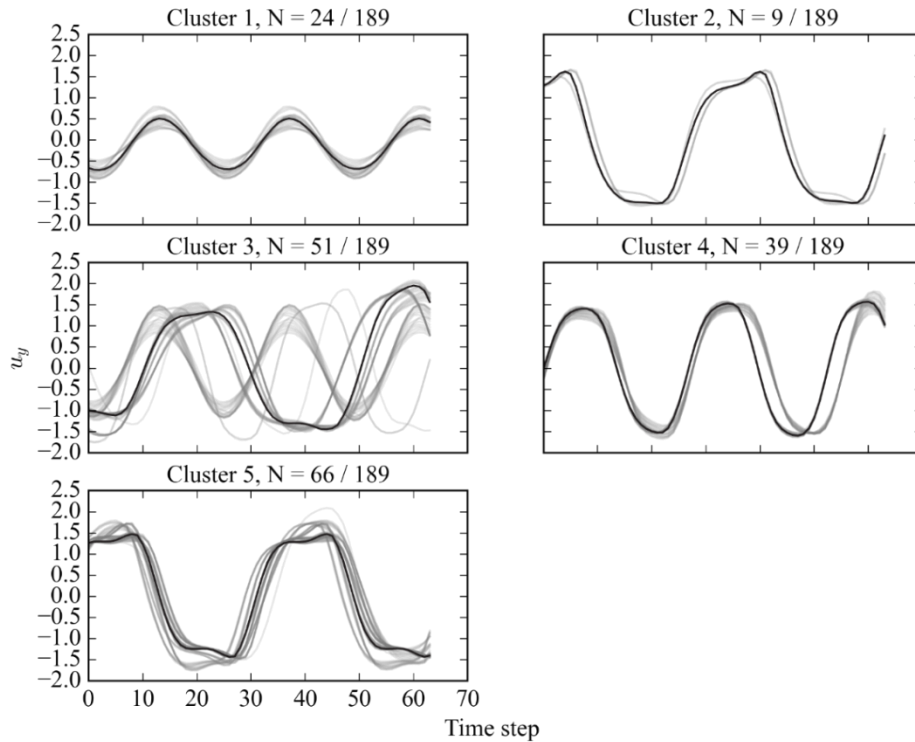


Figure 6.15. Generated clusters using the Hybrid A method at  $Re = 4000$ .

The procedure of clustering based on similarity in shape using dynamic time warping is seen in the samples of cluster 3 in Figure 6.15. The samples in cluster 3 all share relatively similar shapes but are out-of-phase with each other, a product of simply using the splices of the subsequences extracted from the more extended raw time series data. The poor performance of the Dunn index is attributed to the out-of-phase samples in cluster 3, as the calculation of the Dunn index uses the pairwise distances of the samples. The underlying clustering approach was visualized by the two-dimensional latent space of the generated clusters using Hybrid A, shown in Figure 6.16.



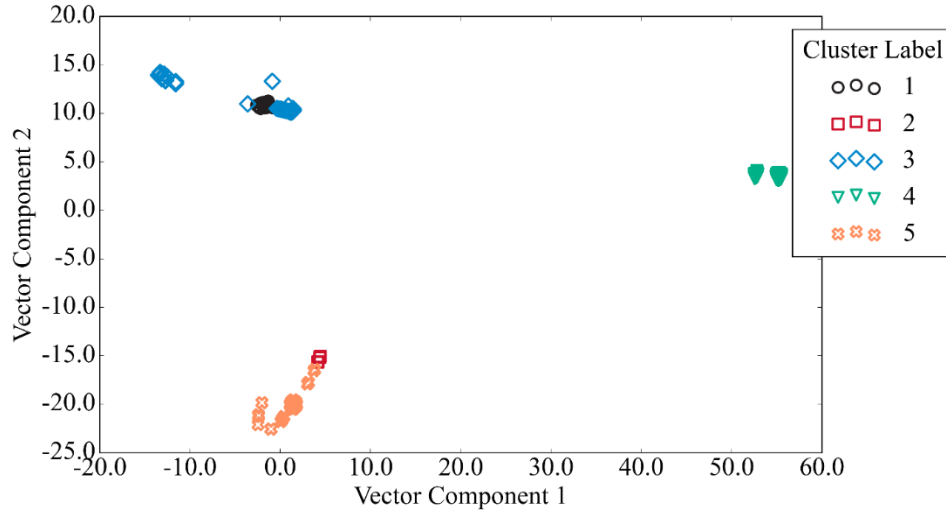


Figure 6.16. Cluster t-SNE distribution using Hybrid A method at  $Re = 4000$ .

The primary and secondary clusters identified at each node and the corresponding percentage of each are summarized in Table 6.8.

Table 6.8: Vortex Shedding Map Cluster Candidates for Hybrid A at  $Re = 4000$

$\lambda^*$	$A^*$	Cluster Candidate		Candidate Proportion [%]	
		Primary	Secondary	Primary	Secondary
2	0.1	4	3	85.7	14.3
2	0.3	3	5	95.2	4.8
2	0.5	5	3	52.4	47.6
4	0.1	4		100	
4	0.7	3	1	85.7	14.3
4	0.9	1		100	
6	0.3	5		100	
6	0.5	5	2	90.5	9.5
6	0.7	5	2	66.7	33.3

Using the primary cluster candidates and the associated weights of the clusters identified at each node, the normalized amplitude and wavelength plane were populated in Figure 6.17.

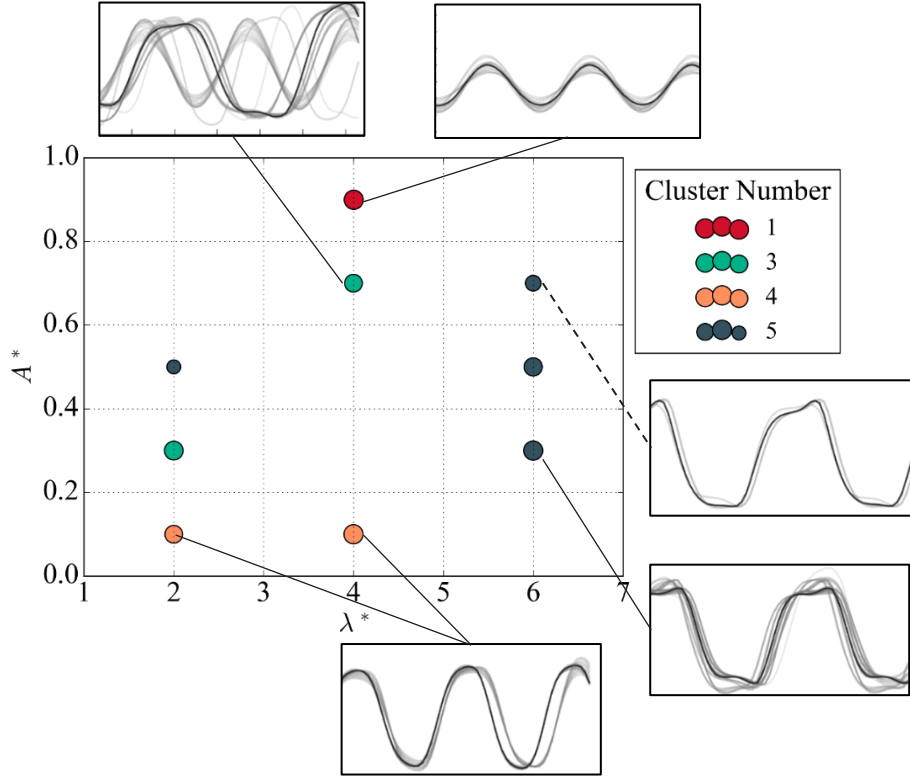


Figure 6.17. Vortex shedding map using Hybrid A method at  $Re = 4000$ .

The vortex shedding map produced identifies several varying regions of vortex shedding behaviour. First, the generated cluster 1 is designated to the node at  $(\lambda^*, A^*) = (6, 0.7)$  which resembles a 2P mode but with a reduced first peak in the pattern. Cluster 2 is primarily for the low-frequency case  $\lambda^* = 2$  with it occurring mainly at  $A^* = 0.3$  and as the secondary mode at  $A^* = 0.5$ . The low amplitude,  $A^* = 0.1$ , space exhibits cluster 3, the regular sinusoidal pattern indicative of 2S behaviour. A similar sinusoidal pattern is observed with cluster 4, differing by a lower observed amplitude in the signal. Finally, cluster 5 is primarily located at  $\lambda^* = 6$  at  $A^* = (0.3, 0.5)$  with an additional split located at  $(\lambda^*, A^*) = (2, 0.5)$ . The additional node shows a rather weak identification due to the switching between the modes of cluster 5 and cluster 2.

### 6.3.2 Hybrid Method B

The clustering analysis results are presented for the corresponding phases using DBSCAN as the pre-clustering phase and agglomerative for the final merging of clusters. The DBSCAN algorithm identified six clusters with three samples considered outliers. A prototype represents each of the six clusters identified in the pre-clustering phase. The prototype of each cluster represents the time series that minimizes the pairwise distances between itself with the other cluster members. The reduced dataset of prototypes is then used for the final clustering stage.

The final clustering is conducted on the prototype dataset to merge similar samples to the desired number of clusters. Dynamic time warping is used for the distance matrix, which is less computationally expensive on the reduced dataset. The distance matrix is then used in the agglomerative algorithm with the tuned complete initialization method. The clustering performance results of both phases are summarized in Table 6.9.

Table 6.9: Clustering Performance Metrics of Hybrid B Method at  $Re = 4000$

Phase	Clustering Algorithm	Number of Clusters	Evaluation Metric	
			Sil	Dunn
1: Pre-Clustering	DBSCAN	6 (3 Noise Points)	0.7185	0.36049
2: Final Clustering	Agglomerative	5	0.7031	0.39836

The DBSCAN algorithm in the pre-clustering phase produced well-separated clusters resulting in a high silhouette index. The final merging step improved the Dunn index marginally at the cost of a slight reduction of silhouette index. The generated clusters merged for the entire dataset are shown in Figure 6.18.

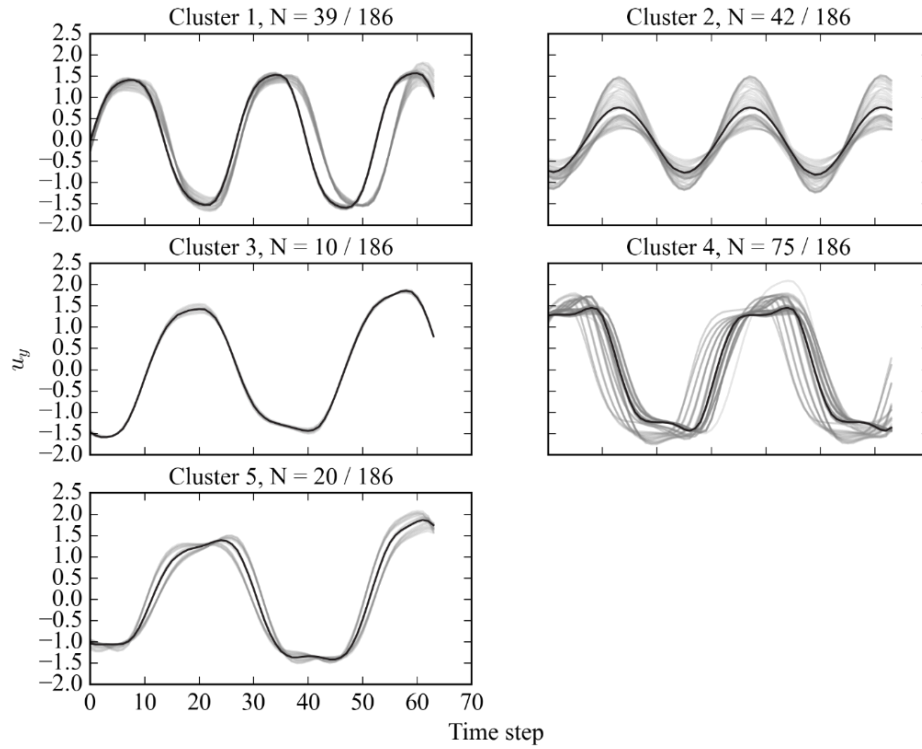


Figure 6.18. Generated clusters by Hybrid B method at  $Re = 4000$ .

The dataset reduction using the  $t$ -SNE method differs from other cases since the DBSCAN algorithm identifies noise points removed from the dataset in the analysis. The latent space for the noise-reduced dataset is shown in Figure 6.19.

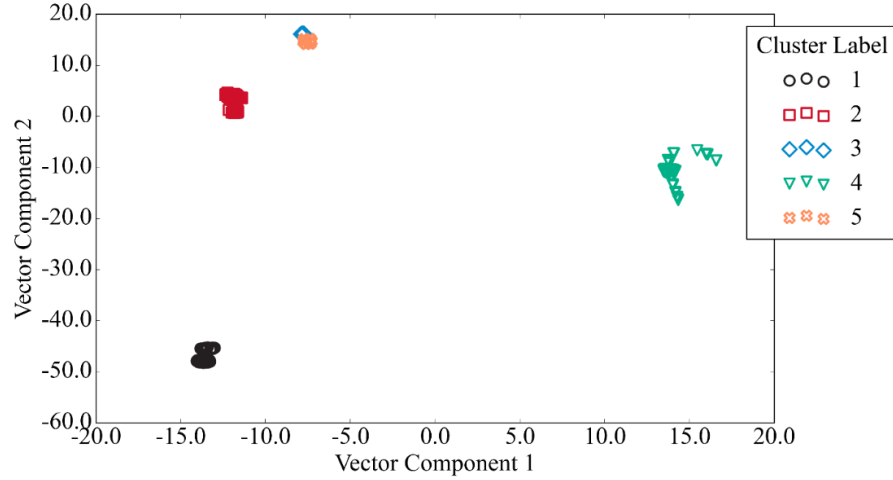


Figure 6.19. Cluster t-SNE distribution using Hybrid B method at  $Re = 4000$ .

The primary and secondary clusters identified at each node and the corresponding percentage of each are summarized in Table 6.10.

Table 6.10: Vortex Shedding Map Cluster Candidates for Hybrid B at  $Re = 4000$

$\lambda^*$	$A^*$	Cluster Candidate		Candidate Proportion [%]	
		Primary	Secondary	Primary	Secondary
2	0.1	1		100	
2	0.3	5	4	95.2	4.8
2	0.5	4	3	52.4	47.6
4	0.1	1		100	
4	0.7	2		100	
4	0.9	2		100	
6	0.3	4		100	
6	0.5	4		100	
6	0.7	4		100	

The associated vortex shedding map was generated based on the primary cluster candidates and the proportions of cluster samples at each node shown in Figure 6.20.

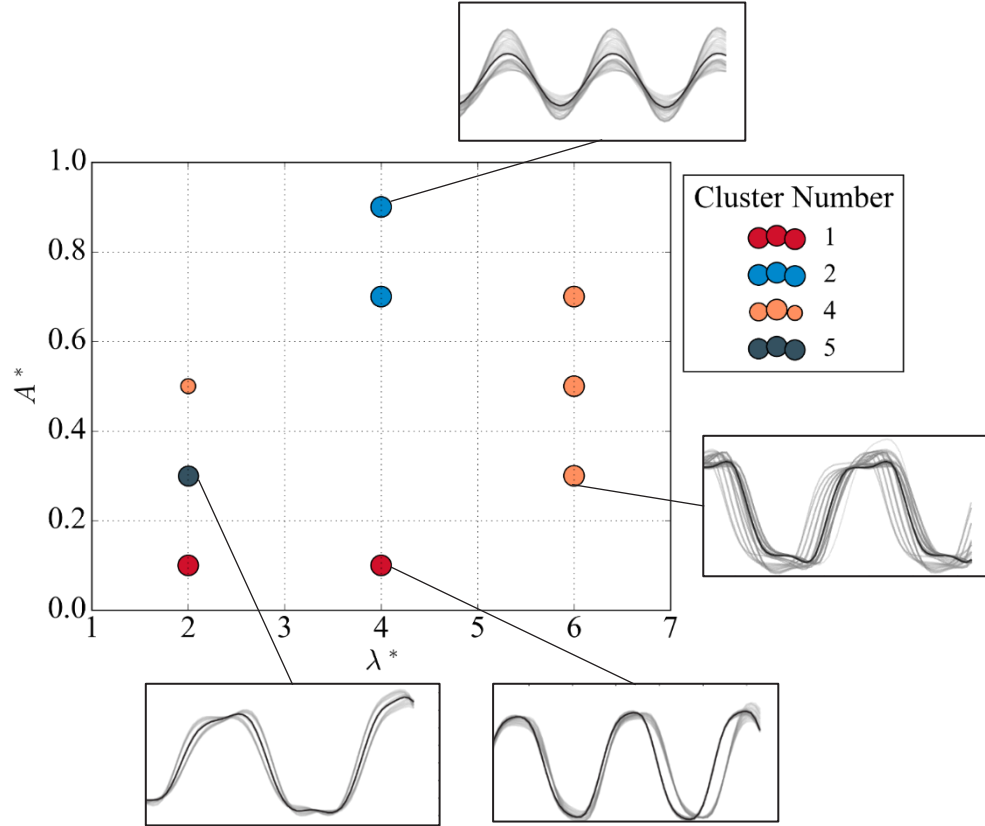


Figure 6.20. Vortex shedding map using Hybrid B method at  $Re = 4000$ .

Using the Hybrid B method, the generated regime map shares its overall structure with the other methods. The nodes along the line  $\lambda^* = 6$  all belong to the primary mode identified in cluster 4 that resembles the 2P mode. The low amplitude nodes along the horizontal line  $A^* = 0.1$  share cluster 1. The sinusoidal mode with lower amplitude is located at  $(\lambda^*, A^*) = (4, 0.7)$  and  $(\lambda^*, A^*) = (6, 0.9)$ . Finally, the node located at  $(\lambda^*, A^*) = (2, 0.5)$  is comprised of two modes labelled by cluster 4 and cluster 3.

### 6.3.3 Hybrid Method C

The pre-clustering and final clustering phase results are presented for the last proposed hybrid methods. The  $k$ -Means algorithm used in the first phase requires the optimum number of clusters to maximize the silhouette index. The evaluation metrics of silhouette and Dunn index for an increasing number of clusters are shown in Figure 6.21.

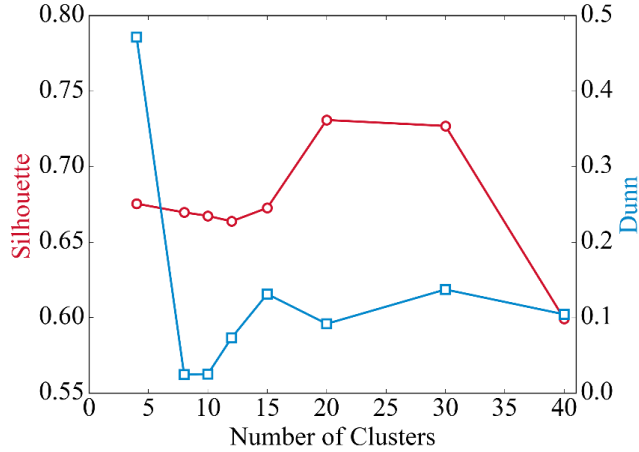


Figure 6.21. Evaluation metrics for the number of clusters generated using *k*-Means.

The optimum number of clusters for the pre-clustering phase was 20 as it maximized the silhouette index representing well-separated clusters. The corresponding clustering performance results of both phases are summarized in Table 6.11.

Table 6.11: Clustering Performance Metrics of Hybrid C Method at  $Re = 4000$

Phase	Clustering Algorithm	Number of Clusters	Evaluation Metric	
			Sil	Dunn
1: Pre-Clustering	k-Means	20	0.7464	0.09175
2: Final Clustering	Agglomerative	5	0.5336	0.11211

The evaluation metrics indicate the quality of clustering in each stage, primarily the significantly separated clusters in the first stage and the more general clusters merged in the final stage. The merging of clusters results in an observed decrease in the silhouette index but improves the performance of the Dunn index. The final cluster performance can be observed in the merging clusters generated by the hybrid method in Figure 6.22.

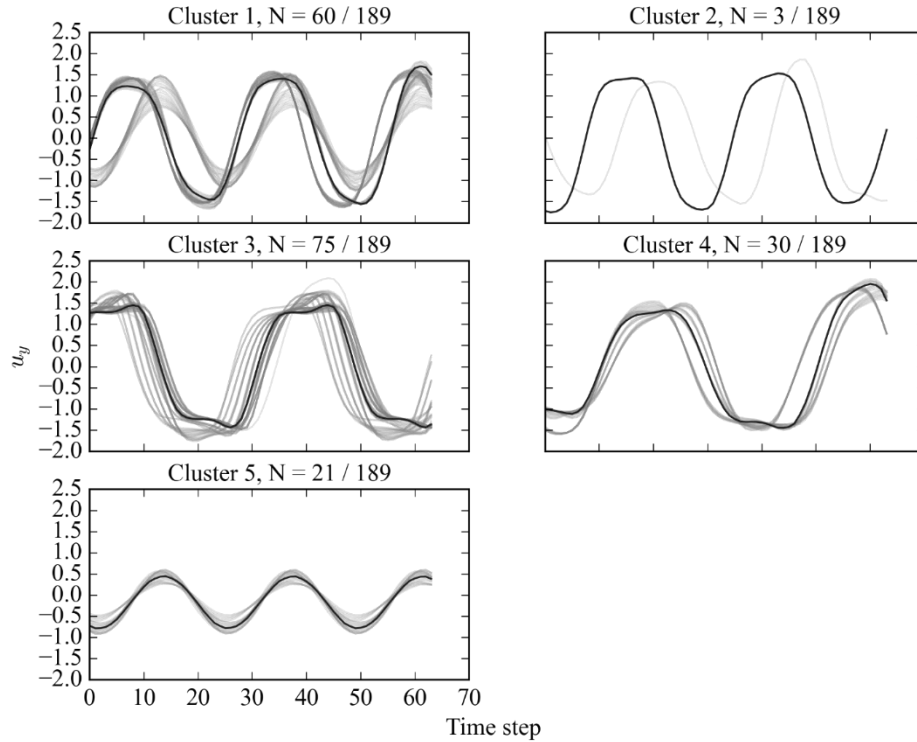


Figure 6.22. Generated clusters by Hybrid C method at  $Re = 4000$ .

The clusters generated overall capture unique and separated subsequence patterns between them. The relatively poor performance in the Dunn index can be identified by the low number of samples identified in cluster 2 and variation in signal in cluster 1. The behaviour of the clustering algorithm in generating the groups can be visualized in the latent space of the dataset converted using  $t$ -SNE shown in Figure 6.23.

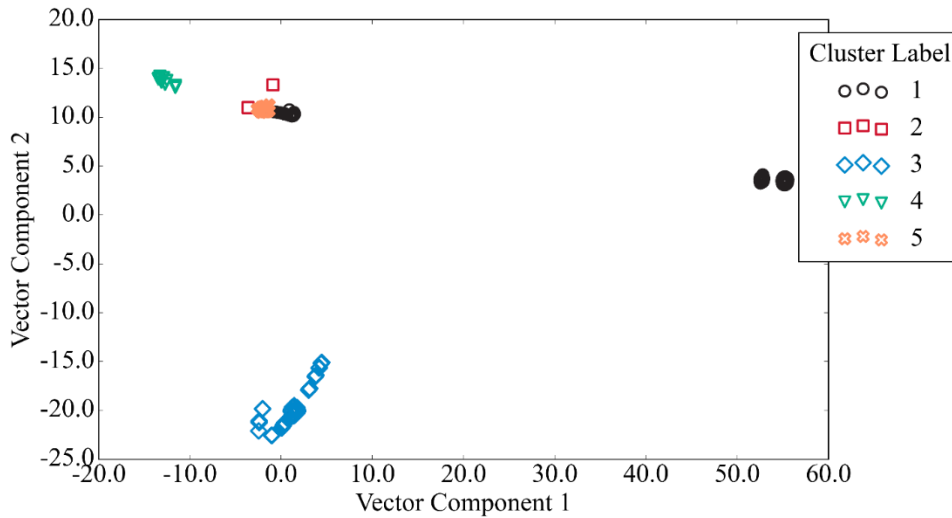


Figure 6.23. Cluster  $t$ -SNE distribution using Hybrid C method at  $Re = 4000$ .

The primary and secondary clusters identified at each node and the corresponding percentage of each are summarized in Table 6.12.

Table 6.12: Vortex Shedding Map Cluster Candidates for Hybrid C at Re = 4000.

$\lambda^*$	$A^*$	Cluster Candidate		Candidate Proportion [%]	
		Primary	Secondary	Primary	Secondary
2	0.1	1	2	85.7	14.3
2	0.3	4	3	95.2	4.8
2	0.5	3	4	52.4	47.6
4	0.1	1		100	
4	0.7	1		100	
4	0.9	5		100	
6	0.3	3		100	
6	0.5	3		100	
6	0.7	3		100	

The associated vortex shedding map was generated based on the primary cluster candidates and the proportions of cluster samples at each node shown in Figure 6.24.

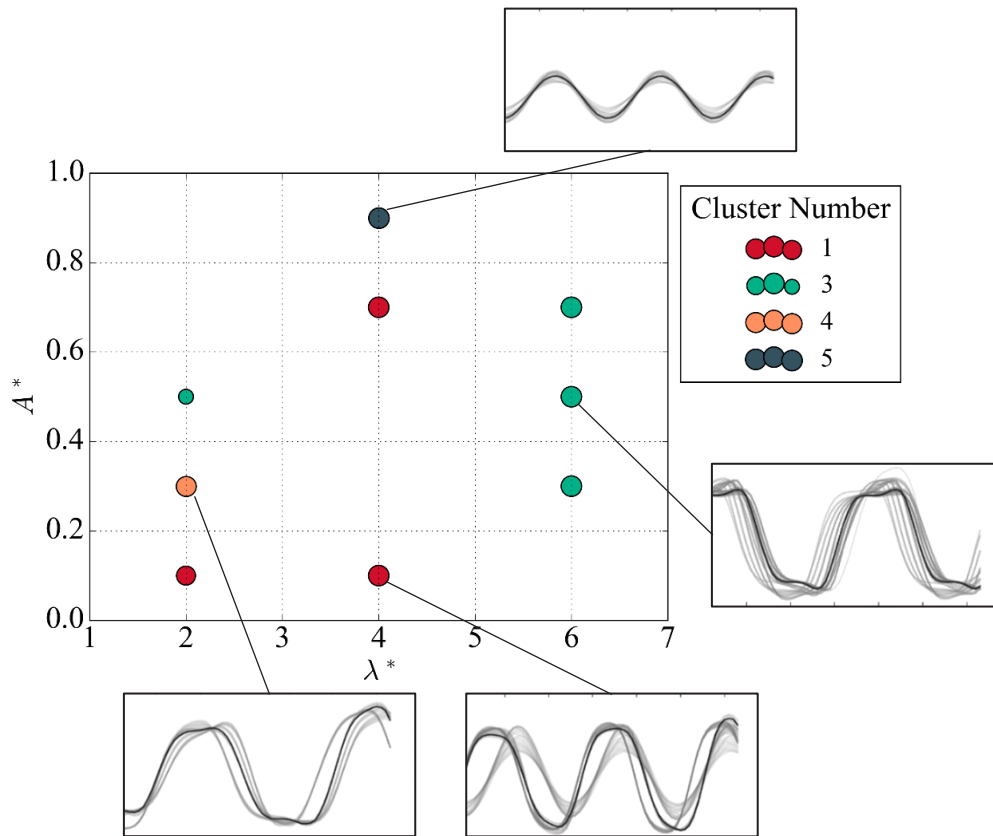


Figure 6.24. Vortex shedding map using Hybrid C method at Re = 4000.



## 6.4 Discussion

The results show that each clustering method offers varying benefits and procedures to the unsupervised clustering task. The respective performance of each method and corresponding qualitative attributes are discussed in this section culminating in an analysis of the vortex shedding

### 6.4.1 Traditional Methods

The partitioning method selected,  $k$ -Means, is a classical clustering method that offers many benefits and limitations for its implementation in this study. A benefit of using partition methods is the relatively simple algorithm implementation, which makes the clusters generated intuitive. The partitioning methods,  $k$ -Means and  $k$ -Medoids, have a linear time complexity  $O(n)$  with the number of data objects,  $n$  [80]. Partitioning methods have limitations that may affect the implementation of these methods in this clustering analysis. First, the nature of the algorithm requires the number of partitions to be specified in the initialization. The number of partitions can be obtained either through domain knowledge or by optimizing cluster numbers. The data-driven goal of this study lends the specification of clusters as a limitation as the required domain knowledge restricts the application of this method for high Reynolds number cases.

The partitioning methods implicitly assume the cluster's shape to be spherical, often visualized as a circle in two dimensions, centred at the mean or medoid. The shape of the clusters can reduce the performance of the clusters as the natural clusters are not guaranteed to be circular. Despite the tuning of the initialization method of the partitioning algorithms, the clustering results can vary between training instances. The initialization of the cluster centers is an integral step in partitioning algorithms, resulting in the method being sensitive to the input data. Furthermore, the mean or medoids of the clusters can be influenced by outliers that have a similar reduction in cluster performance.

Hierarchical methods offer some improvements over the previous partitioning methods. The hierarchical algorithm does not necessarily require the number of partitions specified, as the hierarchy structure can be spliced at the appropriate level to extract the clusters. The hierarchical methods have the shared benefit as partitioning methods that are simple to implement, and the clustering algorithm is relatively intuitive. Furthermore, hierarchical methods have a great visualization power, especially for time series clustering using the generated dendrograms for the generated clusters. Conversely, hierarchical methods have limitations that can affect the clustering performance in this study. The complexity of agglomerative algorithms for hierarchical clustering is considered quadratic,  $O(n^2)$  [38]. The increased computational complexity of hierarchical algorithms restricts the scalability of these methods for larger datasets. Another drawback of hierarchical methods is the sequential approach of merging clusters in the algorithm, which does not allow samples to be reassigned once merged.

Representing the time series data in a reduced dimension can achieve clustering performance gains in many applications because of the reduced complexity and generally more straightforward clustering model. The discrete cosine transform represents the time series in the frequency subspace by the summation of cosine terms of the spectrum. Representing the data by a frequency function may yield benefits for time series cases with increased noise and fluctuations as expected for high Reynolds number cases in this study. The reduced dataset was clustered using the  $k$ -Means method to quantify its effect on clustering.

The benefits and limitations of the selected ordinary clustering methods are summarized in Table 6.13.

Table 6.13: Pros and Cons of Ordinary Clustering Methods

	Pros	Cons
<b>Partitioning Methods (<i>k</i>-Means)</b>	<ul style="list-style-type: none"> <li>• Easy implementation.</li> <li>• Linear time and space complexity.</li> <li>• Reassignment of samples.</li> </ul>	<ul style="list-style-type: none"> <li>• Number of partitions specified.</li> <li>• Sensitive to input data, initial seeds, and outliers.</li> </ul>
<b>Hierarchical Methods (Agglomerative)</b>	<ul style="list-style-type: none"> <li>• Easy Implementation.</li> <li>• Hierarchy dendrogram allows for improved visualization and selection of clusters.</li> </ul>	<ul style="list-style-type: none"> <li>• Cannot reassign erroneous merges of samples.</li> <li>• High time complexity.</li> </ul>

#### 6.4.1.1 Evaluation Metrics

The traditional clustering methods are compared based on the clustering evaluation metrics summarized in Table 6.14.

Table 6.14: Comparison of Ordinary Clustering Methods Based on The Silhouette and Dunn Indices

Type	Clustering Algorithm	Representation Method	Parameters	Evaluation Metric	
				Sil	Dunn
Partitioning	<i>k</i> -Means	Raw Time Series	k-Means++	0.6559	0.15295
Hierarchical	Agglomerative	Raw Time Series	Complete, cosine	0.6794	0.61721
Partitioning	<i>k</i> -Means	Discrete Cosine Transform (DCT)	k-Means++	0.6559	0.15295

The representation method of DCT shows identical clustering performance compared to the *k*-Means method trained using the raw time series data. Although there was no clustering performance improvement over the *k*-Means method trained using the raw time series, the similar clustering validates the use of the DCT as a data reduction method and is selected for its expected gains for the higher Reynolds number cases.

The hierarchical method of the agglomerative algorithm exceeds the performance of the partitioning methods in both a slight improvement of silhouette index and a more dramatic increase in Dunn index. Although high Dunn is desirable for separate and highly compact clusters, the objective of maximizing the Dunn index can result in very few samples being clustered together, which achieved high compactness at the risk of generalizability. The clusters generated by the agglomerative algorithm shown in Figure 6.7 show the limitations of the high Dunn index as Cluster 5 has fewer samples than the other clusters. In this case, the limited number of clusters affects the other clusters as samples of the other clustering methods identified as 2PO are lost in the crowded clusters generated by the agglomerative method. Despite the risk of generalizability and the effect on the vortex shedding map generated, the clustering performance of the agglomerative algorithm is still competitive and is considered for the high Reynolds number case.

### 6.4.1.2 Visual Analysis

The single-step methods using ordinary clustering are validated in the generation of vortex shedding maps by comparing them to the reference map produced by Morse and Williamson [7]. The vortex shedding map produced using the raw  $k$ -Means method is overlaid with the regimes in the reference map as shown in Figure 6.25.

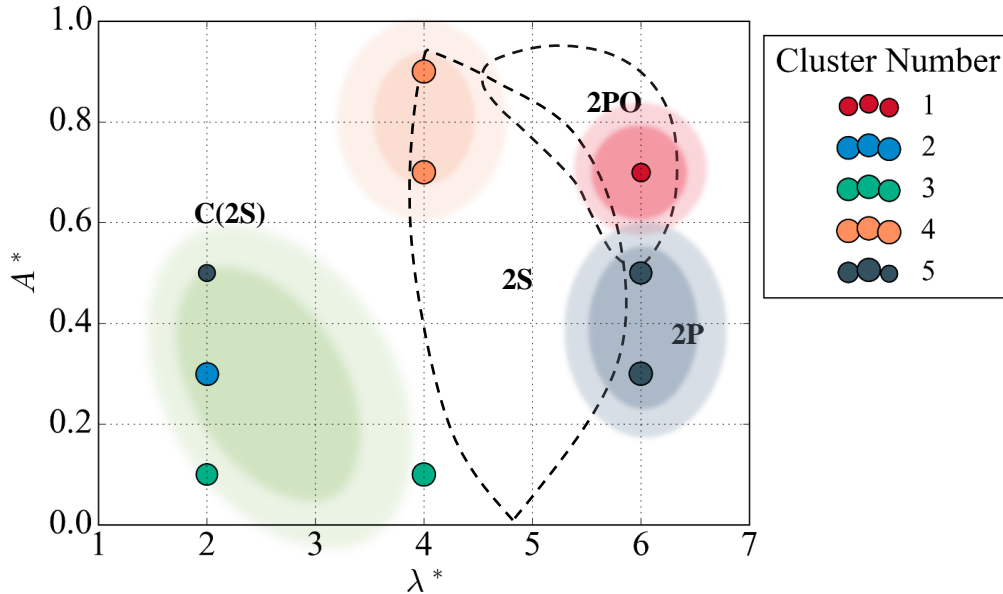


Figure 6.25. Overlaid benchmark regimes on vortex shedding map produced with  $k$ -Means at  $Re = 4000$ .

The overall groups generated by the clustering method fit within the expected regimes produced by Morse and Williamson [7]. The specific case of  $k$ -Means, both with raw and DCT time series, isolate a cluster designated for the 2PO mode. Cluster number 4 and 5 both exclusively inhabit the regions of 2S and 2P modes in the benchmark map. The C(2S) region shows more variation in the assigned cluster numbers. The nodes located in this region at low values of non-dimensional amplitude  $A^* < 0.2$  are strongly controlled by the clear sinusoidal cluster of number 3.

The similarities of the vortex shedding maps produced are apparent in the map generated with the agglomerative algorithm overlaid with the reference map regions, as shown in Figure 6.26.

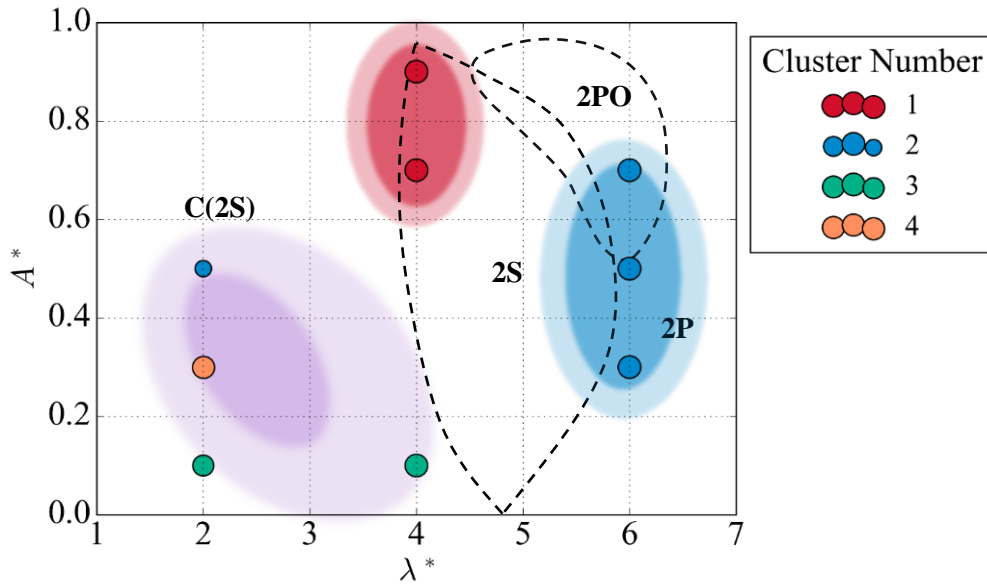


Figure 6.26. Overlaid benchmark regimes on vortex shedding map produced with agglomerative at  $Re = 4000$ .

The general regions identified by cluster 4 and cluster 2 fit within the expected modes of 2P and 2S, respectively. In this case, the cluster in the 2PO region is designated cluster 4, which is not unexpected as the vortex shedding behaviour of 2PO is similar to 2P and can switch intermittently.

#### 6.4.2 Hybrid Methods

The hybrid methods proposed in this thesis all have varying attributes that make them desirable or inherent limitations. Hybrid Method A implements two stages of the partitioning algorithm  $k$ -Medoids. The partitioning method shares the benefits of noted for  $k$ -Means of simple implementation and linear time complexity  $O(n)$  with the number of data objects  $n$  [80]. The partitioning method is limited by the initialization method, which can produce varying clustering results between training instances. The  $k$ -Medoids algorithm is considered an improvement over  $k$ -Means concerning the robustness of the methods. The improved robustness from the  $k$ -Medoids is attributed to the objective function using the median sample as the cluster center. The performance gain of  $k$ -Medoids is debated in the data mining community, especially for cases with increased dimensionality of the centroids, such as the case of this study of time series clusters [81]. As with all partitioning methods, the number of clusters is required in the initialization, which is a limitation in this data-driven study due to the restriction of domain knowledge.

The method implemented in Hybrid B uses DBSCAN for the initial clustering and agglomerative for final clustering. An advantage of DBSCAN is the limited number of parameters required to be initialized, which does not include the number of partitions. The DBSCAN algorithm excels in defining highly separated clusters with unbounded application parameters and identifying outliers. The hierarchical method shares the benefits and limitations previously discussed.

Hybrid method C was specifically designed for optimal clustering performance based on the search of ordinary clustering methods. Independently, both the  $k$ -Means and Agglomerative provide excellent clustering results shown in the analysis of single-stage clustering.

Table 6.15: Pros and Cons of Hybrid Clustering Methods

	<b>Pros</b>	<b>Cons</b>
<b>Hybrid A</b> ( <i>k</i> -Medoids)	<ul style="list-style-type: none"> <li>• Easy implementation.</li> <li>• Linear time and space complexity.</li> <li>• Reassignment of samples.</li> </ul>	<ul style="list-style-type: none"> <li>• Number of partitions specified.</li> <li>• Sensitive to input data, initial seeds, and outliers.</li> </ul>
<b>Hybrid B</b> (DBSCAN, Agglomerative)	<ul style="list-style-type: none"> <li>• Minimal parameters.</li> <li>• Automatic outlier detection.</li> <li>• Arbitrary cluster shapes.</li> <li>• Easy Implementation.</li> <li>• Hierarchy dendrogram allows for improved visualization and selection of clusters.</li> </ul>	<ul style="list-style-type: none"> <li>• Cannot reassign erroneous merges of samples.</li> <li>• High time complexity.</li> </ul>
<b>Hybrid C</b> ( <i>k</i> -Means, Agglomerative)	<ul style="list-style-type: none"> <li>• Easy implementation.</li> <li>• Hierarchy dendrogram allows for improved visualization and selection of clusters.</li> </ul>	<ul style="list-style-type: none"> <li>• Number of partitions specified.</li> <li>• Sensitive to input data, initial seeds, and outliers.</li> </ul>

#### 6.4.2.1 Evaluation Metrics

The hybrid methods are compared based on the clustering evaluation metrics that are summarized in Table 6.16.

Table 6.16: Comparison of Hybrid Clustering Methods Based on The Silhouette and Dunn Indices

<b>Clustering Method</b>	<b>Pre-Clustering Algorithm</b>	<b>Final Clustering Algorithm</b>	<b>Evaluation Metric</b>	
			<b>Sil</b>	<b>Dunn</b>
Hybrid A	<i>k</i> -Medoids	<i>k</i> -Medoids	0.4971	0.01430
Hybrid B	DBSCAN	Agglomerative	0.7031	0.39836
Hybrid C	<i>k</i> -Means	Agglomerative	0.5336	0.11211

The best performing hybrid method concerning silhouette and Dunn index is Hybrid B which has a 31.8% and 255% respective increase over the following best Hybrid C method. The high silhouette index highlights the ability of the DBSCAN method to produce highly separate clusters and reject noise points in the subspace. The Dunn index was marginally improved using agglomeration in the final clustering phase, producing more general clusters. The apparent poor clustering performance of Hybrid A from the low Dunn index is attributed to the use of DTW in the final clustering step. The implementation of DTW with the *k*-Medoids algorithm performed excellently to cluster shape-like patterns that were out of phase with each other. The calculation of the Dunn index uses the pairwise distances of the samples, which results in a low value for the out-of-phase samples.

### 6.4.2.2 Visual Analysis

The hybrid method's ability to generate accurate vortex shedding maps was validated by comparing them to the reference map produced by Morse and Williamson [7]. The vortex shedding map produced using the Hybrid B method is overlaid with the regimes in the reference map, as shown in Figure 6.27.

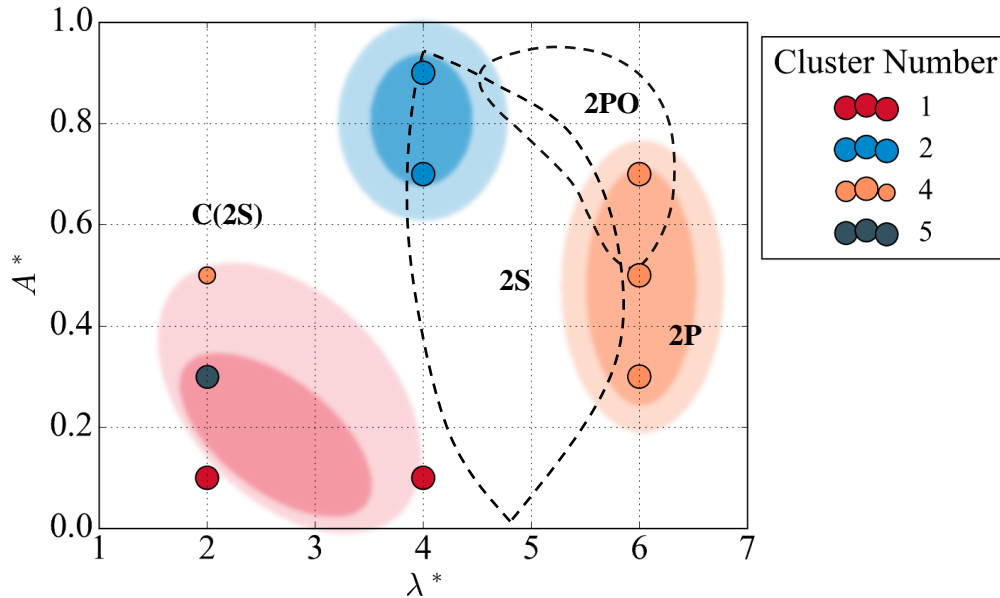


Figure 6.27. Overlaid benchmark regimes on vortex shedding map produced with Hybrid B at  $Re = 4000$ .

The groups of clusters identified by the method overall fit within the expected regions produced by Morse and Williamson [7]. The corresponding pure 2S and 2P regions are inhabited with cluster 2 and cluster 4 points. Although the 2PO mode does not have a distinct cluster identification, the vortex shedding patterns at this node can exhibit both pure 2P signals and intermittent 2PO modes. Similar to all of the clustering methods presented, variation in the groups of clusters in the C(2S) regime is observed. The nodes located in this region at low values of non-dimensional amplitude  $A^* < 0.2$  contains the regular sinusoidal pattern of cluster 1 strictly.

In many of the vortex shedding maps produced, the clusters identified at locations  $(\lambda^*, A^*) = (2, 0.3)$  and  $(\lambda^*, A^*) = (2, 0.5)$  show signals resembling that of the 2P mode, which is unexpected in this region. The seemingly misclassification is attributed to the C(2S) small vortices that coalesce in the near wake, which can corrupt the patterns in the data. Furthermore, the force in phase with acceleration contour graph produced by Morse and Williamson [7] provides more insight on the vortex shedding behaviour at these points, as shown in Figure 6.28.

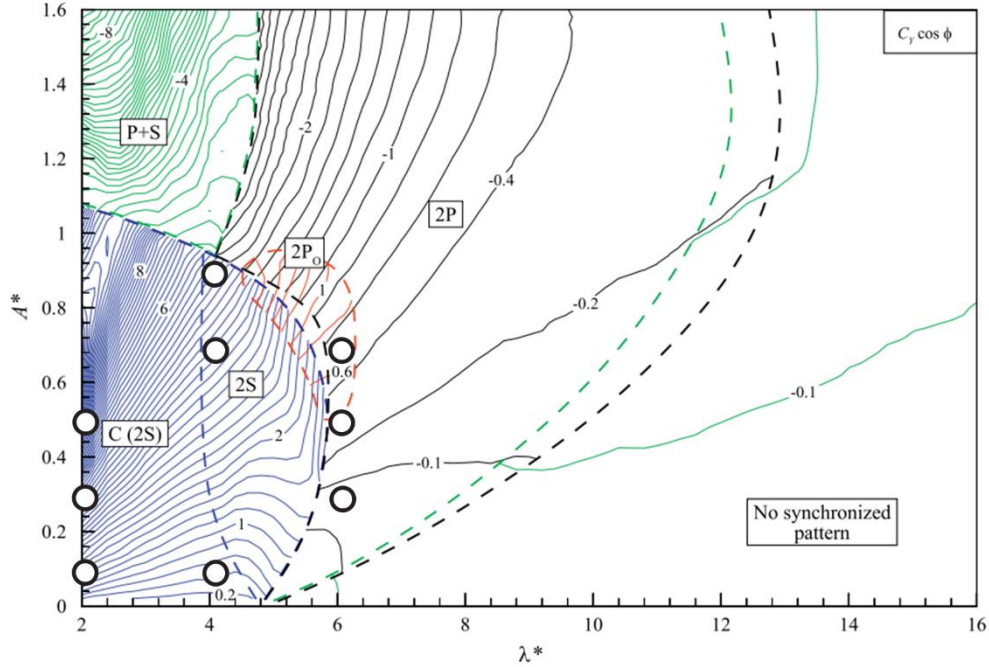


Figure 6.28. Force in phase with acceleration,  $C_y \cos \phi$ , contour graph at  $Re = 4000$  [7].

The two points of interest, specifically the node located at  $(\lambda^*, A^*) = (2, 0.5)$  is located at a point where many force in phase contours merge. The highly positive phase with acceleration may be the cause of the observed varying vortex shedding behaviour seen in this unstable region of the map.

## 6.5 Summary

This chapter presents a combination of unsupervised clustering strategies applied to a benchmark low Reynolds number case of vortex shedding for an oscillating cylinder. The quality of the clustering analysis was evaluated using internal clustering metrics, visual analysis of the clusters, and finally exploring the patterns of the generated vortex shedding maps.

Table 6.17: Final Clustering Performance Metrics of Proposed Methods at  $Re = 4000$

Type	Clustering Algorithm	Representation Method	Evaluation Metric	
			Sil	Dunn
Partitioning	k-Means	Raw Time Series	0.6559	0.15295
Hierarchical	Agglomerative	Raw Time Series	0.6794	0.61721
Partitioning	k-Means	Discrete Cosine Transform (DCT)	0.6559	0.15295
Hybrid	k-Medoids/k-Medoids	Raw Time Series	0.4971	0.01430
Hybrid	DBSCAN/Agglomerative	Raw Time Series	0.7031	0.39836
Hybrid	k-Means/Agglomerative	Raw Time Series	0.5336	0.11211

The results of the data-driven methods for generating vortex shedding maps were determined to agree with the benchmark regime map produced by Morse and Williamson [7]. The regimes maps and corresponding clusters reveal the underlying signatures of each mode based only on the local flow

measurements of the  $y$ -component of velocity. The cluster candidates used to plot the regime map provide additional information on the primary and secondary modes at each node in the parameter space. The additional vortex shedding information in the map provides more precise distinctions between the modes and expected behaviour. The clustering performance and satisfactory agreement with the reference map validate the use of the clustering methods for the application of vortex shedding map generation for an oscillating cylinder at a low Reynolds number.

In conclusion, this chapter addressed the objective of implementing a data-driven approach that requires less input data and supervision through local flow measurements and unsupervised clustering. The presented methods are extended in Chapter 7 to produce vortex shedding maps to an application of high Reynolds numbers with more complex vortex structures that become indistinguishable from traditional methods.



# Chapter 7

## Vortex Shedding Map at High Reynolds Number

In this chapter, the unsupervised clustering methods validated in Chapter 6 for the low Reynolds number case are extended to a higher Reynolds number case to produce a vortex shedding map for more complex flow regimes.

The main contributions of this chapter include gaining insights into the underlying dynamical regimes of the vortex shedding through the map generation. Secondly, quantify the clustering performance to extract meaningful patterns from the local flow field experiencing increased instability and mixing due to the increased dissipation energy of the flow.

### 7.1 Methodology

A similar methodology was implemented in this chapter to study the vortex shedding map generation of the clustering methods. The subsequence from the local flow time series data was extracted using the same method as the low Reynolds number. The clustering analysis was then performed using the ordinary and hybrid methods presented in Chapter 6. The performance of each clustering method was quantified using the internal evaluation metrics and visually validated using the subsequence plots. Additionally, insight on the cluster generation was gained from the distribution of clusters in the  $t$ -SNE embedded latent space. Finally, the vortex shedding maps were generated using the proportion of clusters that decomposed each node into the primary and secondary time series signatures identified by the clustering methods.

#### 7.1.1 Data Exploration

The high Reynolds number dataset [65] sampled the normalized amplitude–wavelength plane of forced oscillations along five sampling lines. The nodes selected in this parameter space for the cluster analysis are shown in Figure 7.1 based on the observed vortex shedding behaviour.

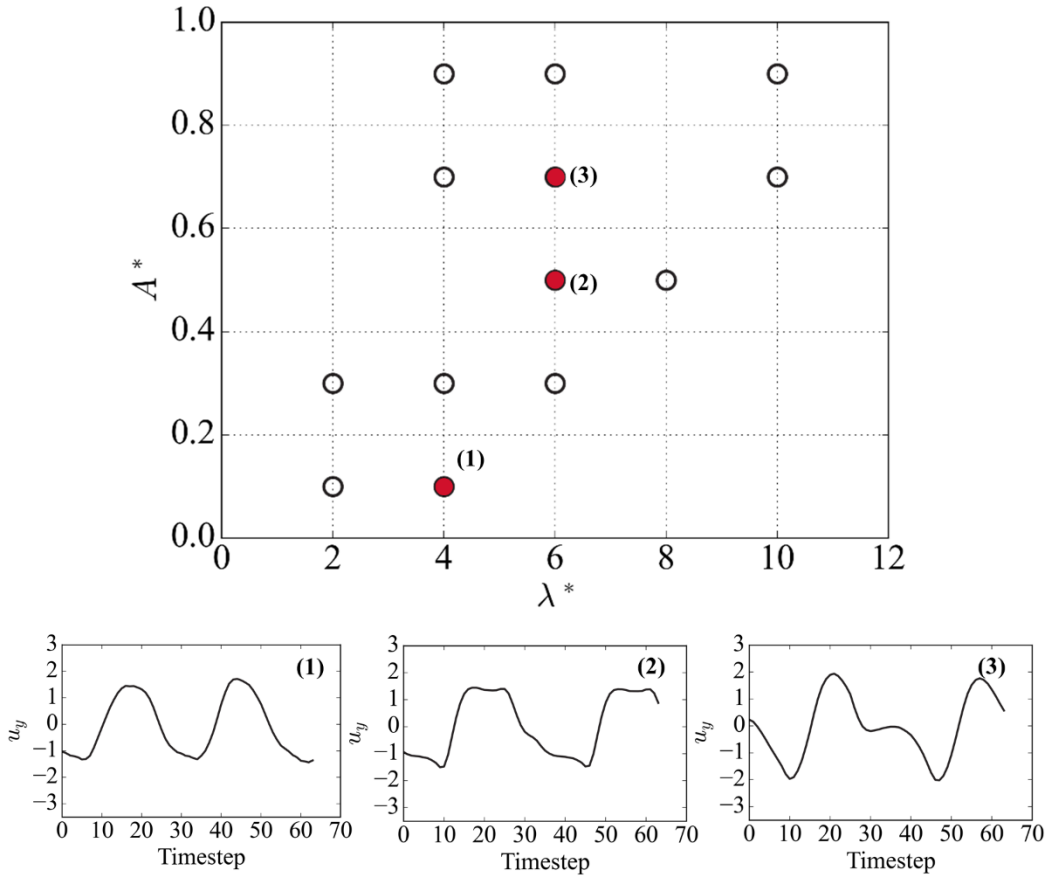


Figure 7.1. Dataset sampled nodes in normalized amplitude–wavelength plane [7].

The clustering dataset included more nodes from the normalized amplitude and wavelength space due to the observed patterns in the local measurements. The repeated patterns in the local flow measurements were then isolated using the subsequence extraction method.

### 7.1.2 Subsequence Extraction

The quality of extracted subsequences for the high Reynolds number cases is pivotal in the clustering results since more fluctuating signals are observed. The window size used for the matrix profile algorithm was confirmed for applying the high Reynolds number case by visual inspection of a relatively consistent pattern and a variable pattern signal shown in Figure 7.2.

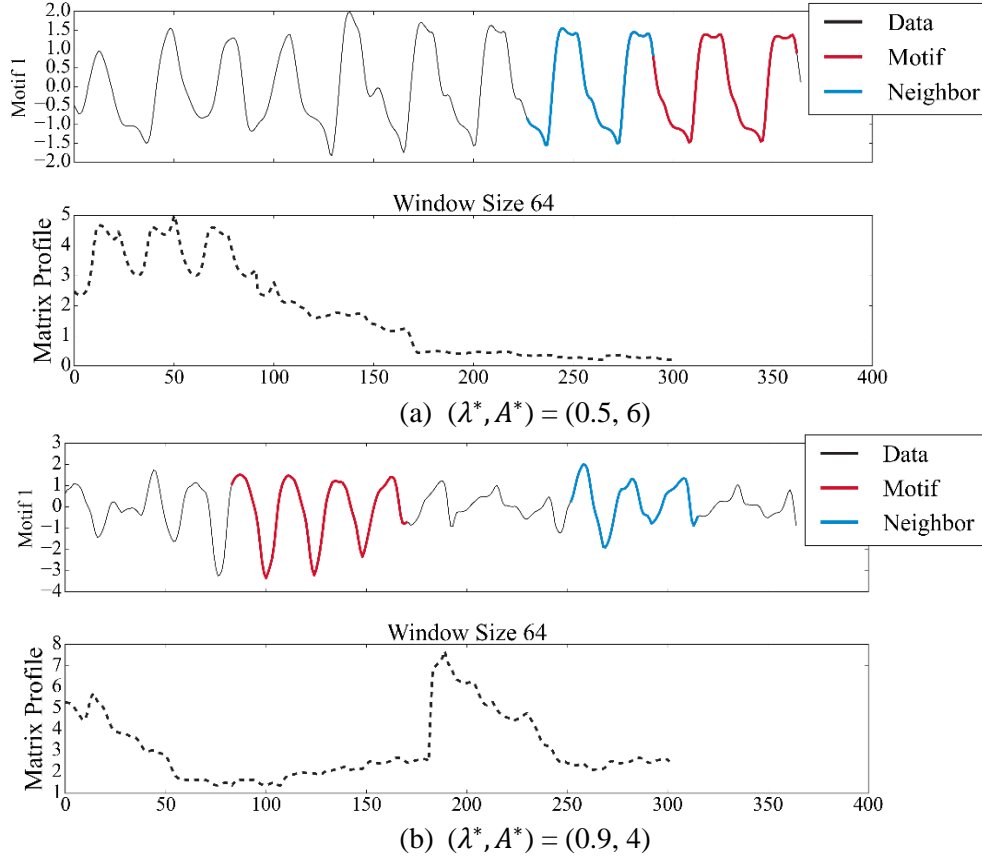


Figure 7.2. Motif extraction for high Reynolds number signals of (a) consistent pattern observed at  $(\lambda^*, A^*) = (0.5, 6)$  and (b) a relatively unstable signal observed at  $(\lambda^*, A^*) = (0.9, 4)$ .

The window size of 64 extracts meaningful patterns from the sample signals even with an unstable signal shown in Figure 7.2 (b). Smaller window sizes would not capture the repeated global patterns, and more oversized windows would not extract valuable patterns to aid in the clustering analysis.

## 7.2 Proposed Traditional Clustering Methods

Applying the clustering methods selected from the low Reynolds number cases requires the number of clusters to be specified. Since this study limits the domain knowledge required for map generation, the optimum number of clusters must be determined. The number of clusters should optimize the clustering performance and the insights that can be extracted through clustering. To ensure that the generated clusters can provide value in the map generation, a limit of 10 clusters is specified. A rule of thumb method of determining the number of clusters is by using the size of the training dataset such that the number of clusters is equal to  $\sqrt{n/2}$  for a dataset of  $n$  instances [34]. This approximation expects to have each cluster with  $\sqrt{2n}$  samples in each. Applying this simple method yields an estimated number of clusters of 11 which qualifies the set limit of 10 clusters for our application.

### 7.2.1 $k$ -Means

The first step in implementing the  $k$ -Means method was determining the number of clusters to consider for this case. The number of clusters selected was nine due to the balanced evaluation metrics of silhouette

and Dunn index. The  $k$ -Means method validated in the previous chapter was implemented using the same initialization method of  $k$ -means++. The clustering performance for the high Reynolds number case was quantified using the internal metrics summarized in Table 7.1.

Table 7.1: Clustering Performance Metrics of  $k$ -Means Method at  $Re = 10,000$

Clustering Algorithm	Evaluation Metric	
	Sil	Dunn
k-Means	0.5750	0.49261

The evaluation metrics of the clusters indicate a good balance between the silhouette and Dunn index, which should result in well-clustered samples. The corresponding clusters generated are shown in Figure 7.3.

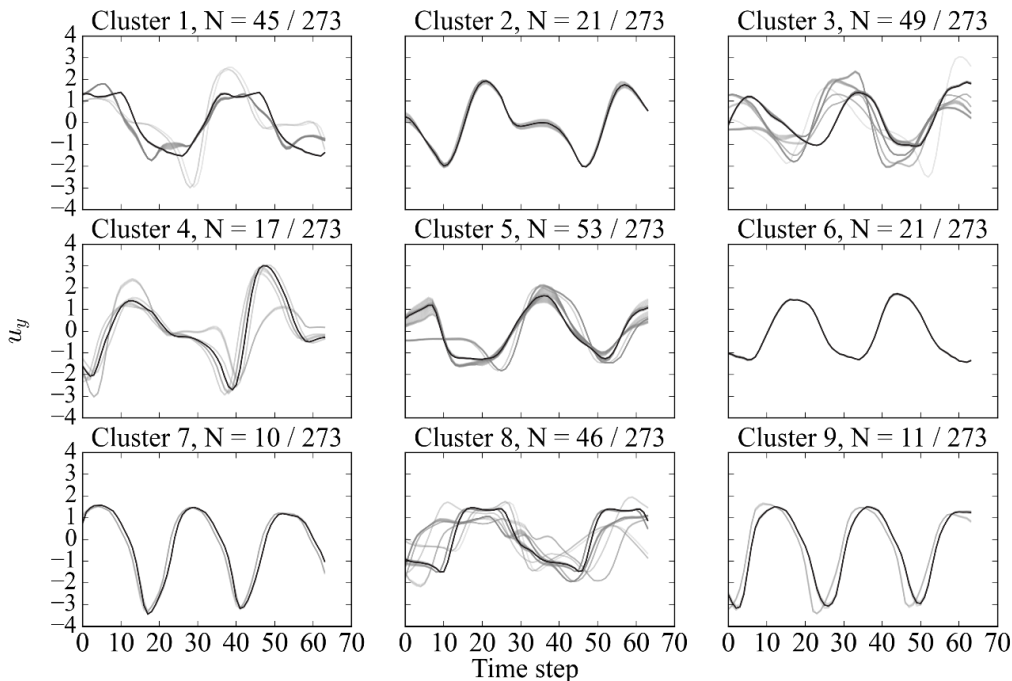


Figure 7.3. Generated clusters by  $k$ -Means method at  $Re= 10,000$ .

From the generated clusters, some patterns begin to appear. Clusters 7 and 9 show a similar pattern of regular sinusoid action with prominent peaks and troughs in the signal. A similar pattern with low amplitude is observed in cluster 6. The expected pattern of the 2PO mode is observed in cluster 2 and cluster 4, highlighted by a smaller peak in-between the peaks. More irregular signals are highlighted in clusters 1, 3, 5, 8. Despite the variations in the signals of clusters 1 and 8, an underlying pattern persists of smaller dual-peaks. Clusters 3 and 5 show a more regular smaller amplitude sinusoid action of the clusters. The cluster distribution in the bi-dimensional embedding latent space generated using  $t$ -SNE is shown in Figure 7.4.

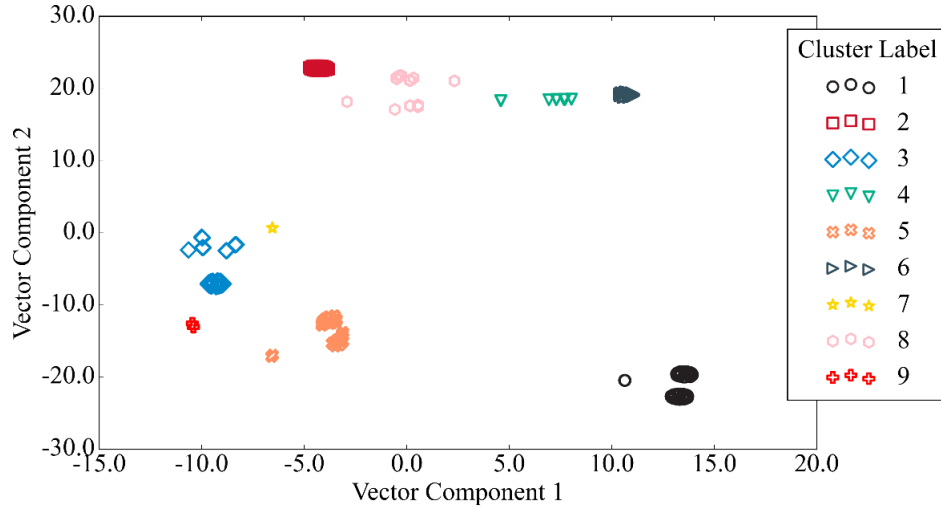


Figure 7.4. Cluster t-SNE distribution at  $Re = 10,000$  using k-Means.

The primary and secondary clusters identified at each node and the corresponding percentage of each are summarized in Table 7.2.

Table 7.2: Vortex Shedding Map Cluster Candidates for  $k$ -Means at  $Re = 10,000$

$\lambda^*$	$A^*$	Cluster Candidate		Candidate Proportion [%]	
		Primary	Secondary	Primary	Secondary
2	0.1	1		100	
2	0.3	1		100	
4	0.1	6		100	
4	0.3	5	3	52.4	47.6
4	0.7	8		100	
4	0.9	9	7	52.4	47.6
6	0.3	3		100	
6	0.5	8		100	
6	0.7	2		100	
6	0.9	4	1	81	14.3
8	0.5	3	8	81	19
10	0.7	5		100	
10	0.9	5		100	

The primary clusters and the relative weight of each label are plotted together on the map shown in Figure 7.5.

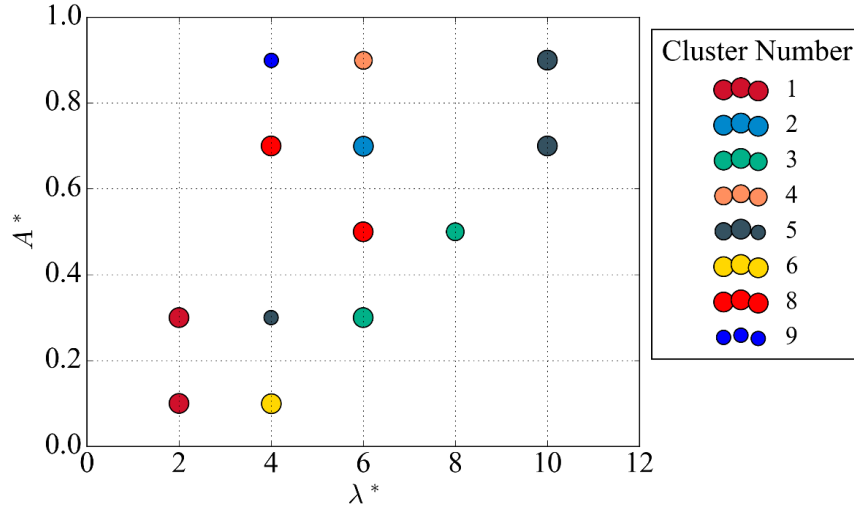


Figure 7.5. Vortex shedding map using  $k$ -Means method at  $Re = 10,000$ .

There are a few groups of clusters in the non-dimensional amplitude and wavelength plane of similar clusters. First, the two nodes located on the  $\lambda^* = 10$  line are both identified by cluster number 5. A similar mode to cluster 5 identified as cluster 3 is shown in the proximity at locations  $(\lambda^*, A^*) = (6, 0.3)$  and  $(8, 0.5)$ . The smaller non-dimensional amplitude values along the line  $\lambda^* = 2$  exhibits the behaviour denoted by cluster 1, which is a more irregular pattern. The similarities of clusters 3 and 5 are shown by the split distribution of the node located at  $(\lambda^*, A^*) = (4, 0.3)$ , which evenly shares the weight of both clusters. The highly regular clusters 7 and 9 are identified at the node  $(\lambda^*, A^*) = (4, 0.9)$  which corresponds to a strong 2S behavior.

### 7.2.2 Agglomerative

The validated agglomerative method was built using the same complete linkage and cosine affinity distance as the method used for the low Reynolds number. The internal indices used to quantify the clustering performance are summarized in Table 7.3.

Table 7.3: Clustering Performance Metrics of Agglomerative Method at  $Re = 10,000$

Clustering Algorithm	Evaluation Metric	
	Sil	Dunn
Agglomerative	0.5760	0.51160

The clusters associated with the evaluation metrics are shown in Figure 7.6

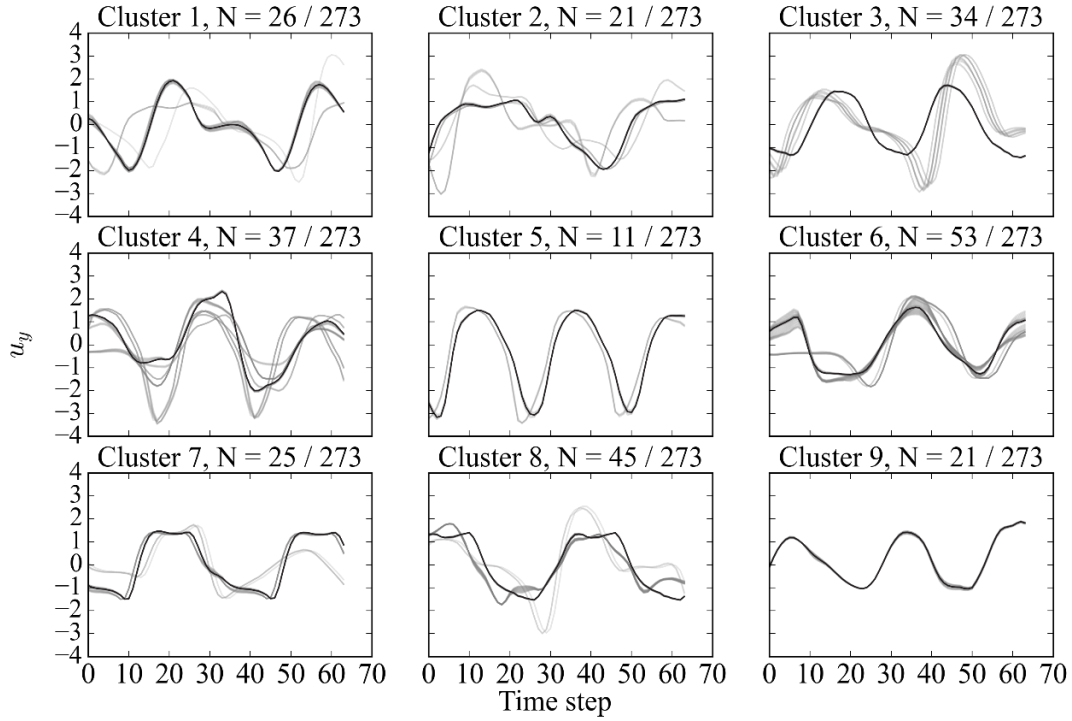


Figure 7.6. Generated clusters by Agglomerative (complete, cosine) method at  $Re = 10,000$ .

The clusters generated using the agglomerative procedure share many similarities to the clusters of the  $k$ -Means method. A sub-peak in-between cycles synonymous with the 2PO mode is identified in clusters 1 and 3. Pure modes were identified in Clusters 1, 3, 5, 7, and 9, with slight variation in the samples. Inconsistent patterns are observed in Clusters 4 and 5, where the clusters would benefit from additionally merging. Additional insight on the clustering procedure can be obtained by the cluster distribution in the two-dimensional  $t$ -SNE latent space shown in Figure 7.7.

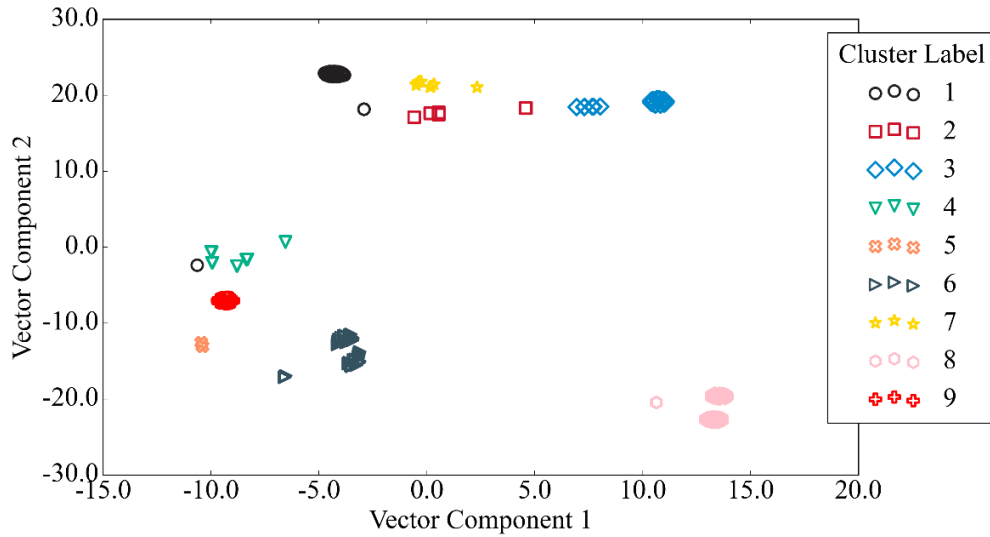


Figure 7.7. Cluster  $t$ -SNE distribution using the agglomerative method at  $Re = 10,000$ .

The clusters of interest, 4 and 5, are relatively close in the latent subspace. Cluster 5 has a relatively low number of samples, and the samples included in cluster 4 seem to be inconsistent with the density of the latent space. The primary and secondary clusters identified at each node and the corresponding percentage of each are summarized in Table 7.4.

Table 7.4: Vortex Shedding Map Cluster Candidates for Agglomerative Re = 4000

$\lambda^*$	$A^*$	Cluster Candidate		Candidate Proportion [%]	
		Primary	Secondary	Primary	Secondary
2	0.1	8		100	
2	0.3	8		100	
4	0.1	3		100	
4	0.3	6	4	52.4	47.6
4	0.7	2	1	81	19
4	0.9	5	4	52.4	47.6
6	0.3	9		100	
6	0.5	7		100	
6	0.7	1		100	
6	0.9	3	2	61.9	19
8	0.5	4	7	81	19
10	0.7	6		100	
10	0.9	6		100	

The vortex shedding map was then plotted with the primary cluster candidates identified, as shown in Figure 7.8.

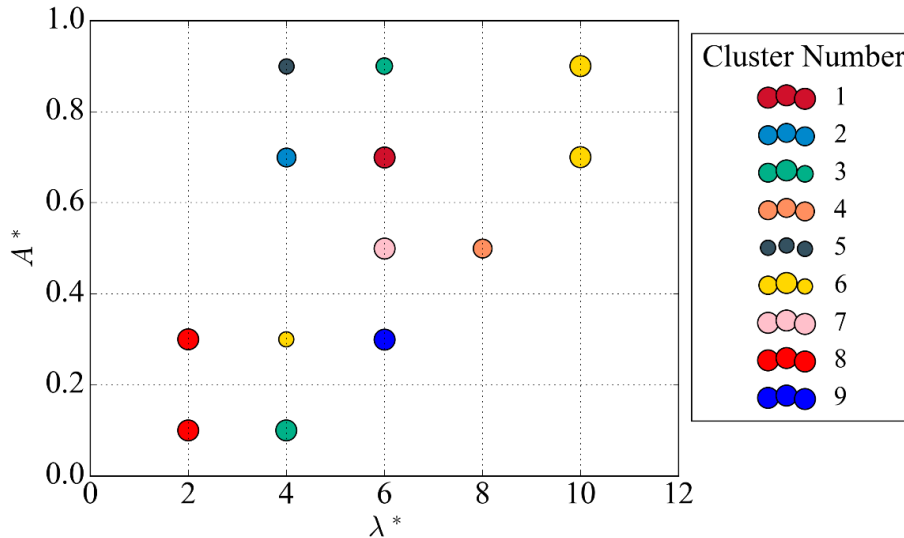


Figure 7.8. Vortex shedding map using the agglomerative method at Re = 10,000.

The areas of interest in this map include the similarly clusters nodes located on the  $\lambda^* = 10$  line, which shares the cluster number 6. The similarities of clusters 4 and 5 are shown by the split distribution of the



node located at  $(\lambda^*, A^*) = (4, 0.9)$ , which evenly shares the weight of both clusters. Cluster number 3 is identified in the parameter space at two unique locations,  $(\lambda^*, A^*) = (6, 0.9)$  and  $(\lambda^*, A^*) = (4, 0.1)$ . The cluster pattern is mainly identified in the lower amplitude case as the cluster shares weight with cluster 2 at the higher amplitude node. These nodes highlight the enhanced vortex shedding behavior identification ability using the primary and secondary cluster candidates and the relative observed weights of each.

### 7.2.3 DCT Time Series Representation with k-Means

The clustering method implemented by representing the time series data using the discrete cosine transform then clustered using  $k$ -Means was selected for this analysis. The clustering performance for the reduced dataset using the discrete cosine transform is summarised in Table 7.5.

Table 7.5: Clustering Performance Metrics of DCT dataset using  $k$ -Means Method at  $Re = 10,000$

Representation Method	Clustering Algorithm	Evaluation Metric	
		Sil	Dunn
DCT	$k$ -Means	0.5938	0.49347

The clusters identified using the  $k$ -Means algorithm trained on the transformed dataset are shown in Figure 7.9.

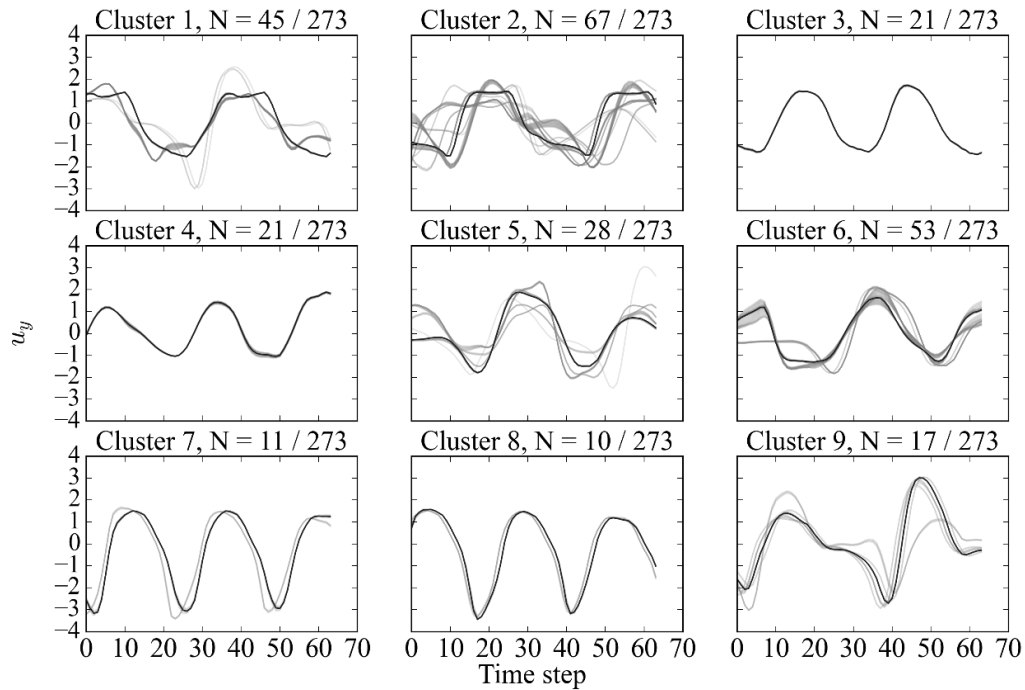


Figure 7.9. Generated clusters by  $k$ -Means method on DCT dataset at  $Re = 10,000$ .

The cluster distribution produced in the two-dimensional latent space generated using  $t$ -SNE is shown in Figure 7.10.

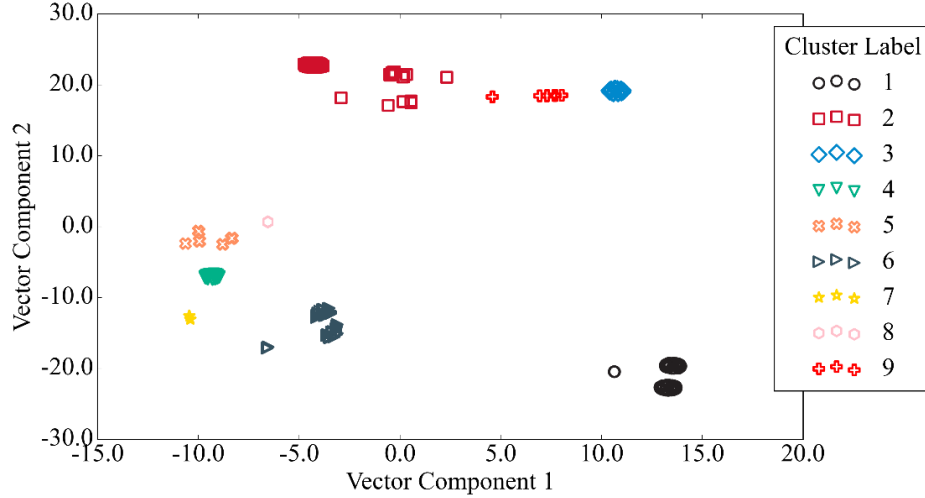


Figure 7.10. Cluster  $t$ -SNE distribution at  $Re = 10,000$  using  $k$ -Means on DCT dataset.

The primary and secondary clusters identified at each node and the corresponding percentage of each are summarized in Table 7.6.

Table 7.6: Vortex Shedding Map Cluster Candidates for  $k$ -Mean on DCT dataset  $Re = 10,000$

$\lambda^*$	$A^*$	Cluster Candidate		Candidate Proportion [%]	
		Primary	Secondary	Primary	Secondary
2	0.1	1		100	
2	0.3	1		100	
4	0.1	3		100	
4	0.3	6	5	52.4	47.6
4	0.7	2		100	
4	0.9	7	8	52.4	47.6
6	0.3	4		100	
6	0.5	2		100	
6	0.7	2		100	
6	0.9	9	1	81	14.3
8	0.5	5	2	81	19
10	0.7	6		100	
10	0.9	6		100	

The corresponding vortex shedding map produced with the primary cluster candidates and the nodes in the non-dimensional amplitude and wavelength plane is shown in Figure 7.11.

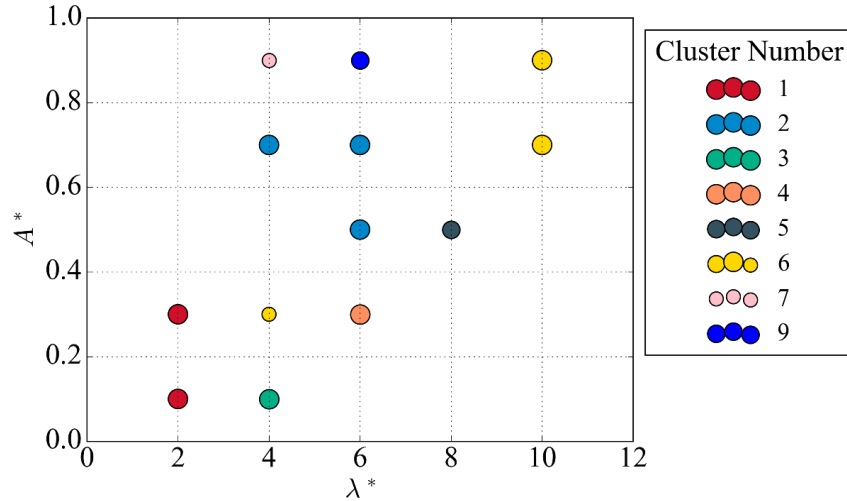


Figure 7.11. Vortex shedding map using  $k$ -Means method on DCT dataset at  $Re = 10,000$ .

The map produced using the DCT and  $k$ -Means method shares many identified clusters in the  $k$ -Means method trained using the raw time series.

### 7.3 Proposed Hybrid Clustering Methods

The three hybrid methods validated in Chapter 6 are compared based on internal evaluation metrics, cluster plots, latent space cluster distribution, and generated vortex shedding map.

#### 7.3.1 Hybrid Method A

The first step in implementing the hybrid method is to determine the number of clusters that produce the first stage's optimum performance. The evaluation metrics of silhouette and Dunn index for an increasing number of clusters are shown in Figure 7.12

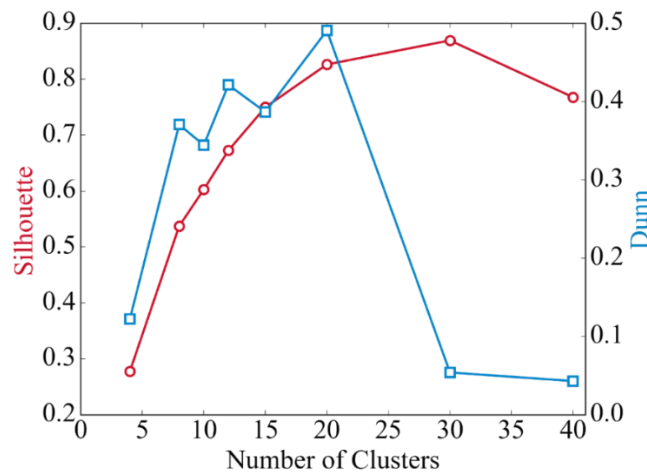


Figure 7.12. Evaluation metrics for the number of clusters generated using  $k$ -Medoids.

The optimum number of clusters was determined to be 30 since the multi-step approach maximizes the separation of the clusters for the first stage. The clustering performance results of both phases are summarized in Table 7.7.

Table 7.7: Clustering Performance Metrics of Hybrid A Method at  $Re = 10,000$

Phase	Clustering Algorithm	Number of Clusters	Evaluation Metric	
			Sil	Dunn
1: Pre-Clustering	k-Medoids	30	0.8624	0.04957
2: Final Clustering	k-Medoids	6	0.3675	0.19489

The generated clusters using the Hybrid A method in the final clustering phase are shown in Figure 7.13

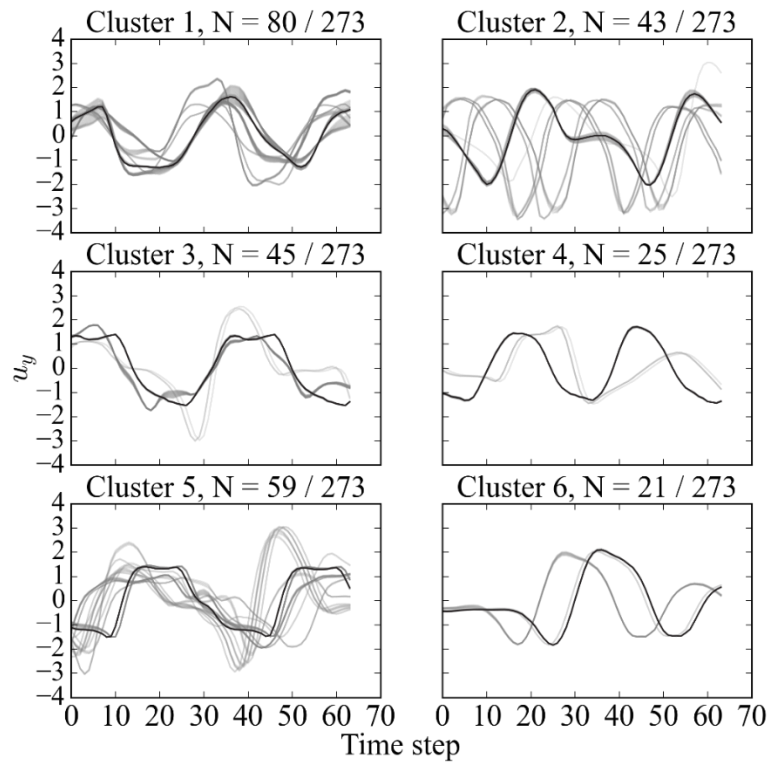


Figure 7.13. Generated clusters using the Hybrid A method at  $Re = 10,000$ .

Visually inspecting the generated clusters, the hybrid method misses clustered samples. The intermediate peak pattern is lost in the clusters, the samples being included in both cluster 2 and cluster 5. The samples included in Cluster 5 show minimal consistent patterns associated uniquely with the cluster. The inability of the clustering method to produce separate and compact clusters is demonstrated with the low Reynolds number in Table 7.7. The basis of the clustering procedure can be additionally investigated using the two-dimensional latent space of the generated clusters using Hybrid A is shown in Figure 7.14.

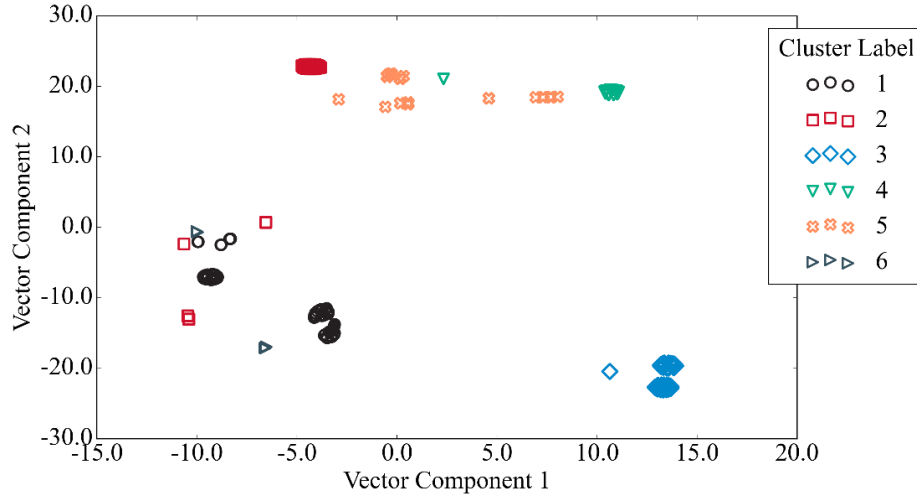


Figure 7.14. Cluster t-SNE distribution using Hybrid A method at  $Re = 10,000$ .

The poor clustering performance, specifically for cluster 5, is shown in the latent space in Figure 7.14. The samples included as cluster 5 span a large region with a relatively low density of points. As the underlying patterns differ, other clustering methods split up this subspace with multiple clusters. The primary and secondary clusters identified at each node and the corresponding percentage of each are summarized in Table 7.8.

Table 7.8: Vortex Shedding Map Cluster Candidates for Hybrid A at  $Re = 10,000$

$\lambda^*$	$A^*$	Cluster Candidate		Candidate Proportion [%]	
		Primary	Secondary	Primary	Secondary
2	0.1	3		100	
2	0.3	3		100	
4	0.1	4		100	
4	0.3	6		100	
4	0.7	5		100	
4	0.9	2		100	
6	0.3	1		100	
6	0.5	5		100	
6	0.7	2		100	
6	0.9	5	3	81	14.3
8	0.5	1	4	81	19
10	0.7	1		100	
10	0.9	1		100	

The vortex shedding map produced by the primary cluster candidates and the corresponding weights is shown in Figure 7.15

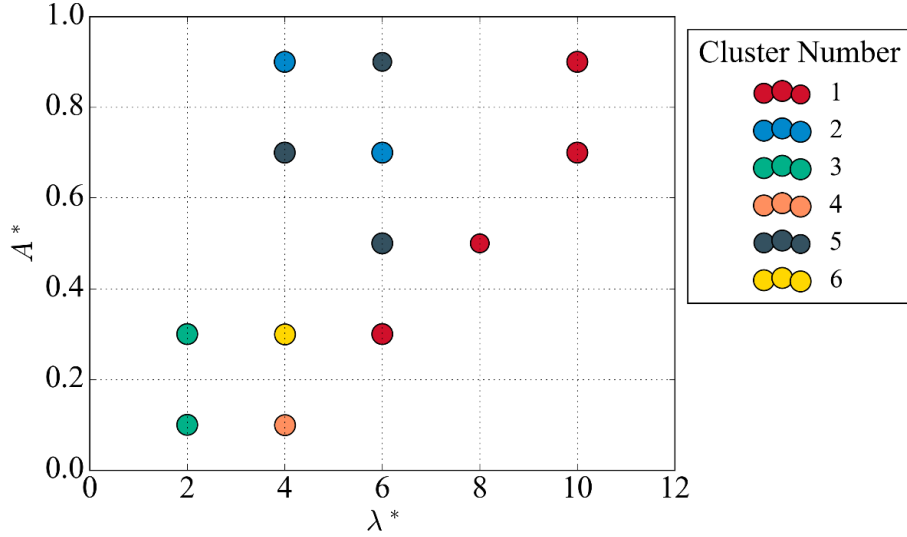


Figure 7.15. Vortex shedding map using Hybrid A method at  $Re = 10,000$ .

The map produced by the Hybrid A method produces very distinct groups of similar clusters. The region along  $\lambda^* = 6, 8, 10$  all exhibit the cluster 1 behaviour of dual peak signals. The region at higher amplitudes and  $\lambda^* = 4, 6$  contain the signatures of clusters 2 and 5. Finally, the lower amplitude and wavelength nodes, including  $\lambda^* < 5$  and  $A^* < 0.4$  contains the cluster numbers 3, 4, and 6.

### 7.3.2 Hybrid Method B

The number of clusters in the pre-clustering phase is not required to be determined since the DBSCAN algorithm automatically finds the optimum number and corresponding noise points. The algorithm found 14 separate clusters and 30 noise points. The clustering performance results of both phases are summarized in Table 7.9.

Table 7.9: Clustering Performance Metrics of Hybrid B Method at  $Re = 10,000$

Phase	Clustering Algorithm	Number of Clusters	Evaluation Metric	
			Sil	Dunn
1: Pre-Clustering	DBSCAN	14 (30 Noise Points)	0.7250	0.20279
2: Final Clustering	Agglomerative	6	0.4822	0.31561

The generated clusters merged for the entire dataset are shown in Figure 7.16.

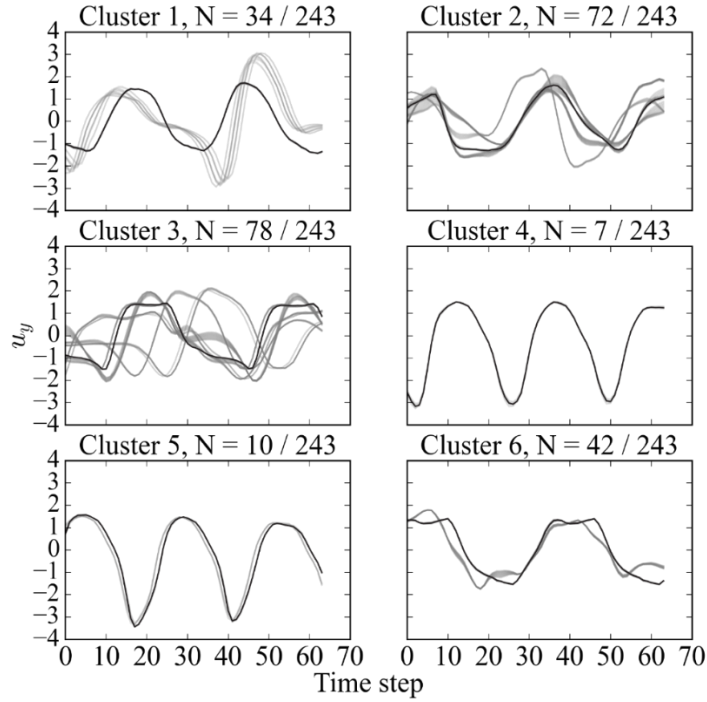


Figure 7.16. Generated clusters by Hybrid B method at  $Re = 10,000$ .

The cluster samples in each label demonstrated the hybrid method B's ability to extract similar shape patterns. Cluster 3 contains out-of-phase samples, but the similarity in shape is observed between the patterns. The dataset reduction using the  $t$ -SNE method differs from other cases since the DBSCAN algorithm identifies noise points that were removed from the dataset in the analysis. The latent space for the noise-reduced dataset is shown in Figure 7.17.

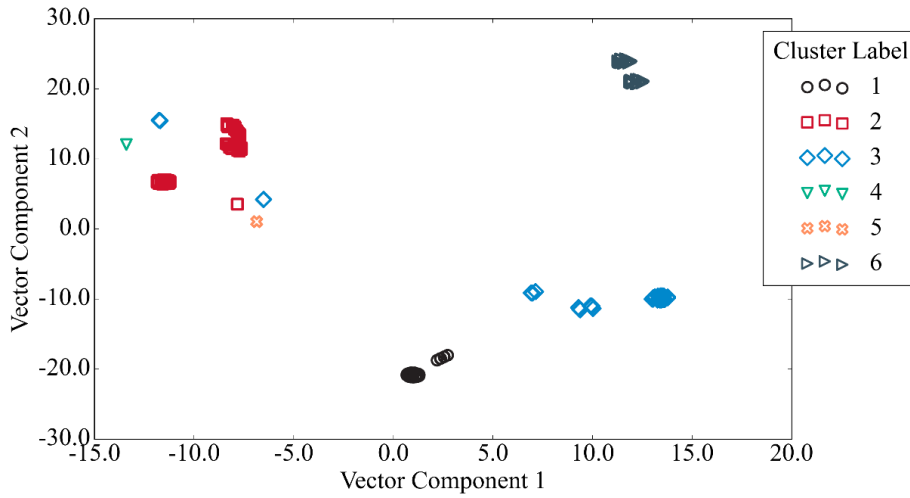


Figure 7.17. Cluster  $t$ -SNE distribution using Hybrid B method at  $Re = 10,000$ .

The primary and secondary clusters identified at each node and the corresponding percentage of each are summarized in Table 7.10.

Table 7.10: Vortex Shedding Map Cluster Candidates for Hybrid B at  $Re = 10,000$

$\lambda^*$	$A^*$	Cluster Candidate		Candidate Proportion [%]	
		Primary	Secondary	Primary	Secondary
2	0.1	6		100	
2	0.3	6		100	
4	0.1	1		100	
4	0.3	3		100	
4	0.7	3		100	
4	0.9	5	4	58.8	41.2
6	0.3	2		100	
6	0.5	3		100	
6	0.7	3		100	
6	0.9	1		100	
8	0.5	2		100	
10	0.7	2		100	
10	0.9	2		100	

The associated vortex shedding map was generated based on the primary cluster candidates and the proportions of cluster samples at each node shown in Figure 7.18.

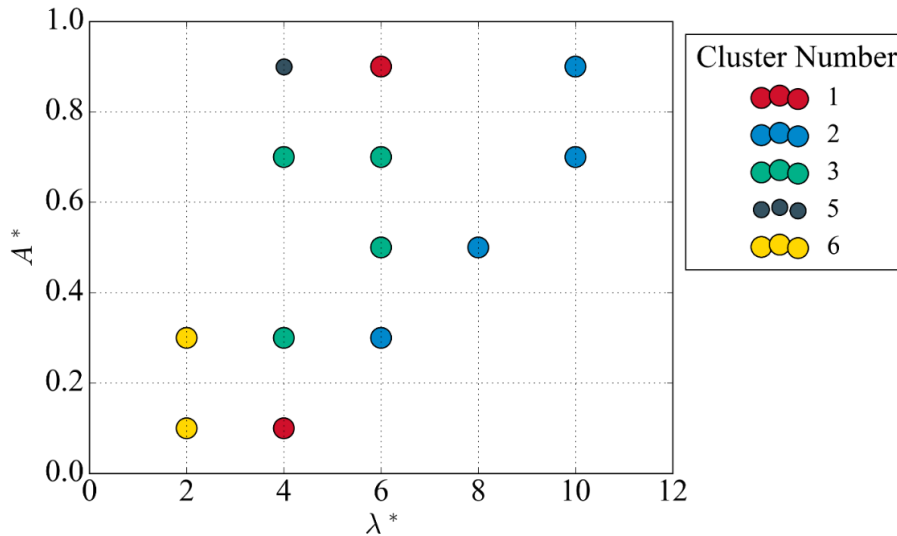


Figure 7.18. Vortex shedding map using Hybrid B method at  $Re = 10,000$ .

The map produced using the Hybrid B method highlights distinct regions of similar cluster results. A regime of cluster number 2 is located along  $\lambda^* = 6, 8, 10$ , which exhibits a dual peak signal behaviour. The middle of the graph contains the members of cluster 3, which contains the signal resembling the 2PO mode with the smaller secondary peak.



### 7.3.3 Hybrid Method C

The pre-clustering phase requires the number of clusters which was determined by the varying evaluation metrics for increasing clusters shown in Figure 7.19.

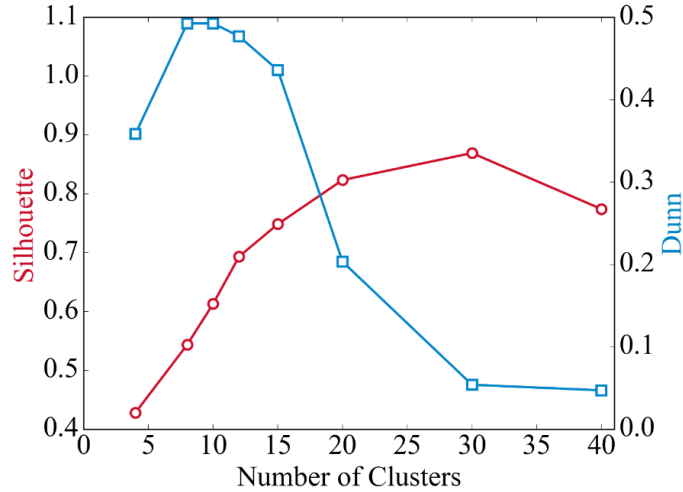


Figure 7.19. Evaluation metrics for the number of clusters generated using  $k$ -Means.

The optimum number of clusters was determined to be 30 since the multi-step approach maximizes the separation of the clusters for the first stage. The clustering performance results of both phases are summarized in Table 7.11.

Table 7.11: Clustering Performance Metrics of Hybrid B Method at  $Re = 10,000$

Phase	Clustering Algorithm	Number of Clusters	Evaluation Metric	
			Sil	Dunn
1: Pre-Clustering	k-Means	30	0.8693	0.05395
2: Final Clustering	Agglomerative	9	0.4694	0.28585

In the final cluster phase, the merged cluster labels produced the subsequence clusters shown in Figure 7.20.

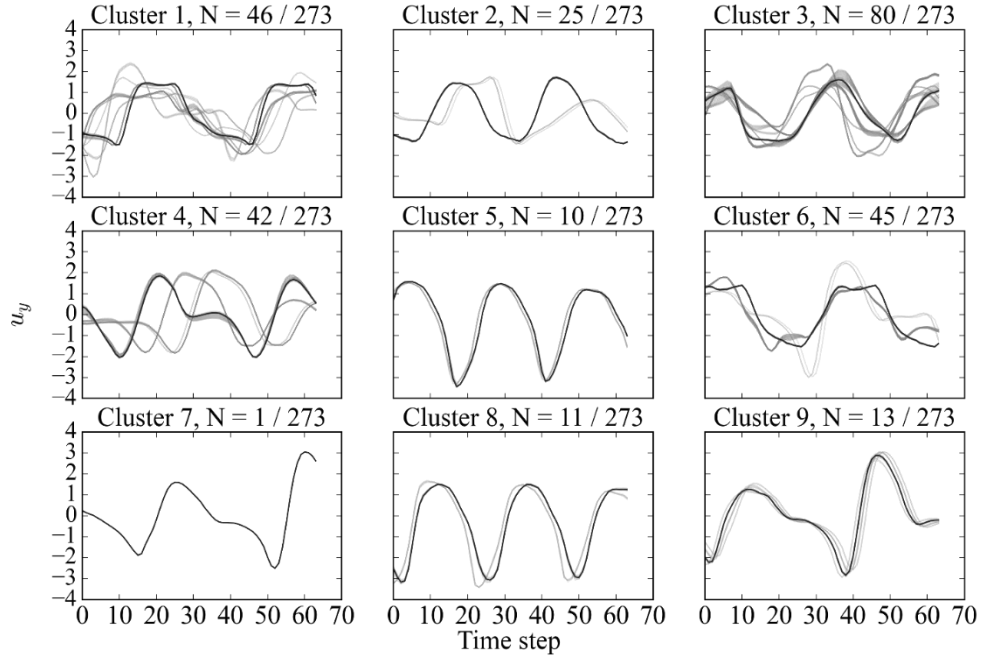


Figure 7.20. Generated clusters by Hybrid C method at  $Re = 10,000$ .

The patterns of interest in the generated clusters include the resemblance of a 2S mode for clusters 5 and 8. Clusters 1, 3, and 6 include more irregular signals, but an underlying pattern of smaller dual peaks can be distinguished. Finally, cluster numbers 4, 7, and 9 exhibits strong 2PO behaviour. The distribution of clusters generated from the algorithm can be visualized represented using the latent space embedding from *t*-SNE shown in Figure 7.21.

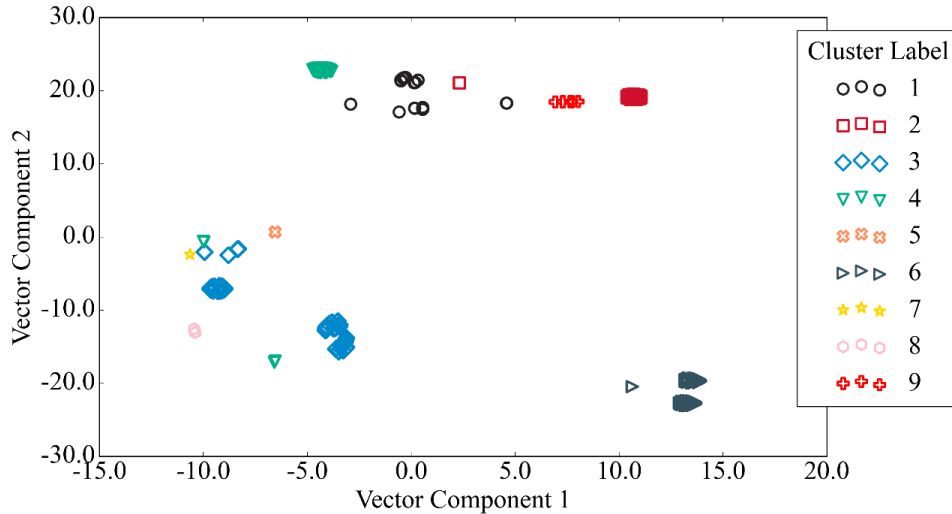


Figure 7.21. Cluster *t*-SNE distribution using Hybrid C method at  $Re = 10,000$ .

The primary and secondary clusters identified at each node and the corresponding percentage of each are summarized in Table 7.12.

Table 7.12: Vortex Shedding Map Cluster Candidates for Hybrid C at Re = 10,000

$\lambda^*$	$A^*$	Cluster Candidate		Candidate Proportion [%]	
		Primary	Secondary	Primary	Secondary
2	0.1	6		100	
2	0.3	6		100	
4	0.1	2		100	
4	0.3	4		100	
4	0.7	1		100	
4	0.9	8	5	52.4	47.6
6	0.3	3		100	
6	0.5	1		100	
6	0.7	4		100	
6	0.9	9	1	61.9	19
8	0.5	3	2	81	19
10	0.7	3		100	
10	0.9	3		100	

The associated vortex shedding map was generated based on the primary cluster candidates and the proportions of cluster samples at each node shown in Figure 7.22.

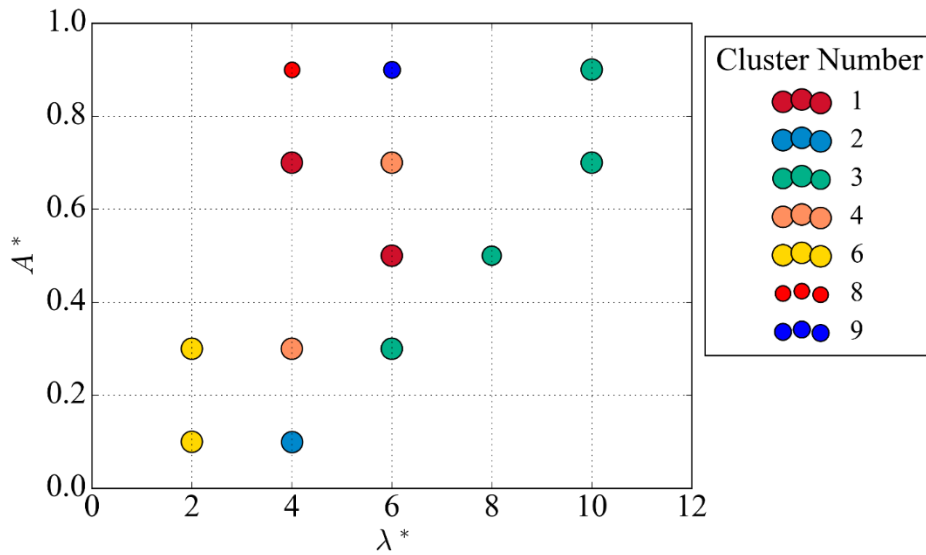


Figure 7.22. Vortex shedding map using Hybrid C method at Re = 10,000.

The map generated can be used to find areas of similar cluster behaviour. Specifically, we see a similar region to the other hybrid methods along  $\lambda^* = 6, 8, 10$ , which exhibits a dual peak signal behaviour.

## 7.4 Discussion

The selected clustering methods validated in Chapter 6 were implemented for the case of a high Reynolds number. The clustering method's ability to produce separate and compact clusters was determined by the internal evaluation metrics summarised in Table 7.13.

Table 7.13: Final Clustering Performance Metrics of Proposed Methods at  $Re = 10,000$

Type	Clustering Algorithm	Representation Method	Evaluation Metric	
			Sil	Dunn
Partitioning	<i>k</i> -Means	Raw Time Series	0.5750	0.49261
Hierarchical	Agglomerative	Raw Time Series	0.5760	0.51160
Partitioning	<i>k</i> -Means	Discrete Cosine Transform (DCT)	0.5938	0.49347
Hybrid	<i>k</i> -Medoids/ <i>k</i> -Medoids	Raw Time Series	0.3675	0.19489
Hybrid	DBSCAN/Agglomerative	Raw Time Series	0.4822	0.31561
Hybrid	<i>k</i> -Means/Agglomerative	Raw Time Series	0.4694	0.28585

The ordinary clustering methods outperformed the hybrid methods regarding the evaluation metrics of silhouette and Dunn index. The reduced clustering performance of the hybrid methods is attributed to the use of dynamic time warping (DTW) in the final clustering phase, which groups signals based on shape and not on time. The pairwise distance calculation in the silhouette and Dunn index will score time series poorly if the patterns are out of phase, even if the shape is consistent in the cluster.

Generally, the clusters produced using the hybrid methods are more similar based on shape than the ordinary methods. The similarity in shape of the hybrid methods will yield better vortex shedding maps based on signature shapes. The improved vortex shedding maps are observed in the plotted domain of non-dimensional amplitude and wavelength. The maps generated using the ordinary methods are more sporadic and require an increased number of clusters to summarize the vortex shedding behaviour.

Despite the lower overall evaluation metrics, it was determined that hybrid methods outperformed single-stage methods in the quality of the clusters based on similarity of shape and the lower number of clusters required to represent the data. Investigating the hybrid methods further, the clusters of Hybrid A were the poorest performing, and many samples were deemed incorrectly clustered in the generated labels. Specifically, the samples of clusters 2 and 5 showed minimal consistent patterns associated uniquely with the cluster. Furthermore, the *t*-SNE latent space demonstrated the limited clustering performance with large regions of relatively low density being clustered. The poor performance is attributed to the use of the *k*-Medoids method in both stages of the hybrid method. The limitations of the *k*-Medoids method were discussed in Chapter 6 with its sensitivity to input data, initial seeds, and outliers. The combination of the same clustering method compounded the limitations and provided no additional benefit in the performance.

The various regions of vortex shedding behaviour at high Reynolds numbers can be obtained by analyzing the produced cluster maps. The non-dimensional amplitude and wavelength plane populated with the corresponding clusters derived using Hybrid Method B is shown in Figure 7.23.

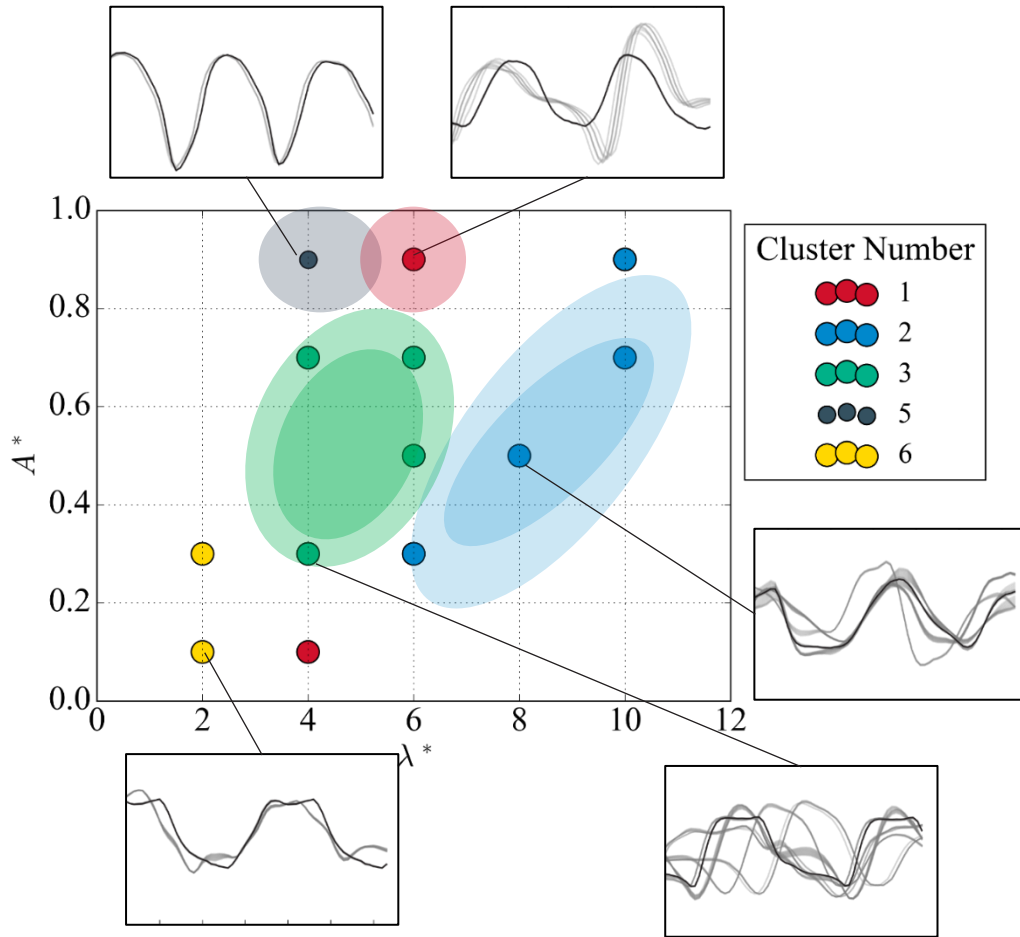


Figure 7.23. Vortex shedding map regions using the Hybrid B method at  $Re = 10,000$ .

The coloured regions provide a qualitative visualization of similar vortex shedding behaviour. The two-level shaded coloured regions are used to proxy the confidence levels of the expected cluster numbers in each zone. Two independent regions are located at the top of the map  $(\lambda^*, A^*) = (4, 0.9)$  and  $(\lambda^*, A^*) = (6, 0.9)$  with consistent vortex shedding signals. The former node signal is periodic with distinct peaks that resemble the 2S mode. The second node signal, denoted with cluster number 1, indicates the 2PO mode with a subpeak in-between the relative peaks generated by the weaker vortex structure. The samples in cluster 1 include samples closer to that of cluster 5 located at the previous node, which indicates a level of overlap between the vortex structures. A larger group of nodes with the same cluster number was identified in the middle of the map,  $3 < \lambda^* < 7$  and  $0.2 < A^* < 0.8$ . Cluster 3 follows differing variations of 2PO and 2P of dual peak and sub-peaks in oscillating actions. The region defined by cluster 2 resembles the signal of the P+S mode determined from the smoothed dual-peak of the oppositely spinning vortices of the P mode and the single peak of the S mode passing the sensor. The clusters identified at the nodes along  $\lambda^* = 2$  and node  $(\lambda^*, A^*) = (4, 0.1)$  are not consistent with a global region of clusters and the field flow behaviour is expected to be chaotic.

The cluster map produced using hybrid method C also outlines various regions of similar fluid flow behaviour. The non-dimensional amplitude and wavelength plane populated with the corresponding clusters derived using Hybrid Method C is shown in Figure 7.24.

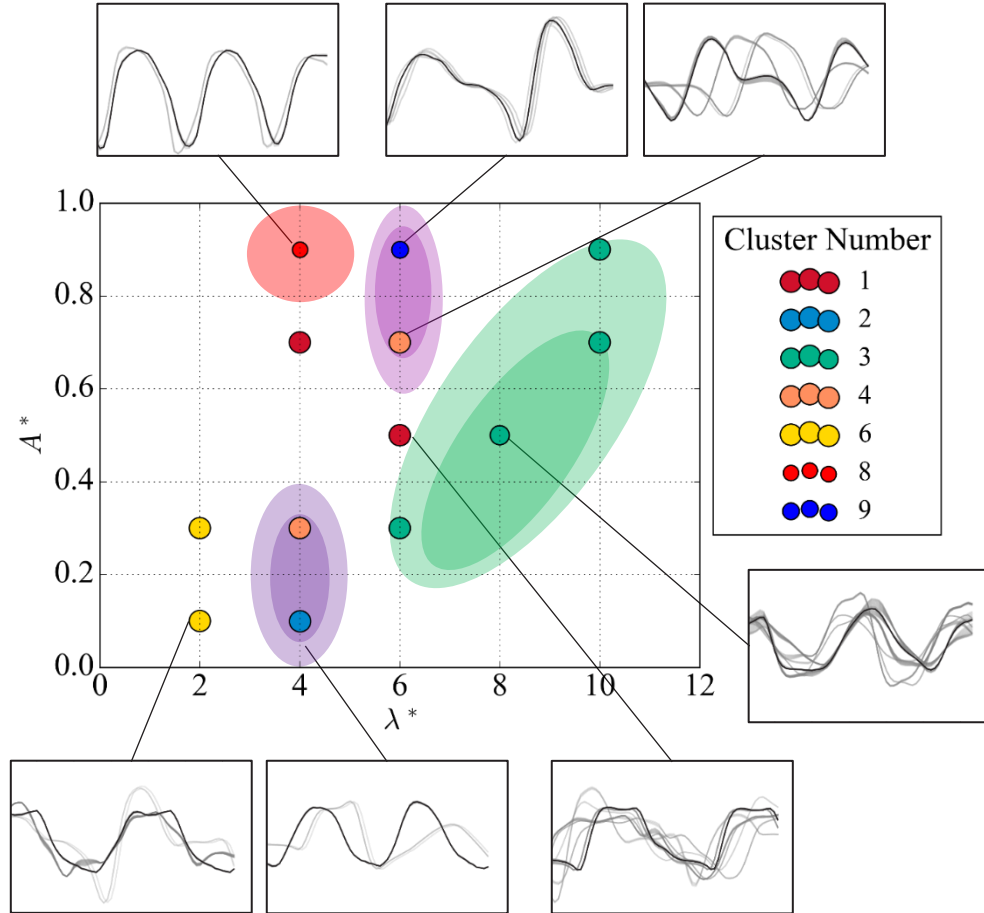


Figure 7.24. Vortex shedding map regions using the Hybrid C method at  $Re = 10,000$ .

The coherent vortex shedding patterns identified in the map first includes the 2S behaviour at the node  $(\lambda^*, A^*) = (4, 0.9)$ . Nearby, two strong 2PO clusters were identified at  $(\lambda^*, A^*) = (6, 0.9)$  and  $(\lambda^*, A^*) = (4, 0.7)$  demonstrated by a weaker vortex being shed in between the oscillations. The former point of cluster 9 shares the similarity in shape but at high peak amplitudes compared to cluster 4. The largest region of similar behaviour is identified by cluster 3, which dominates the higher wavelength portion of the subspace. The signals in cluster 3 demonstrate the behaviour of 2P but appear to have less defined twin peaks. Finally, a region in low amplitude nodes was identified resembling the 2S mode at lower amplitudes compared to that of node  $(\lambda^*, A^*) = (4, 0.9)$ .

From a flow physics perspective, the vortex shedding maps at high Reynolds numbers provide novel insights into the underlying vortex structure interactions. The higher Reynolds number and associated higher flow energy seemed to produce more variation signals and a dissipation effect. The dissipation effect is a product of effectively less viscous flow, which increases flow mixing and creates a more homogenous flow. The signals appeared smooth, with samples with twin peaks not as defined as the low

Reynolds number case at 4000. Furthermore, the high Reynolds number was observed to create more noise in the vortex shedding map with an increased number of nodes with no visible patterns due to the coalescence of vortices. Overall, the signals of the high Reynolds number case were of the larger amplitude of the  $y$ -component of velocity measurement. Specifically, the 2S behaviour showed large amplitudes between peaks in the signals, which aided in identifying these clusters.

The high Reynolds case has clear regions of the 2PO mode demonstrated by the signals of clusters 4 and 9, which show the two pairs of vortices being shed per cycle with one vortex in each oscillation much weaker. The transition mode is in close proximity to the observed P+S signal behaviour of cluster 3 demonstrated by the dual peak of the P mode being separated with a single peak from the S mode. The source of the P+S mode may be attributed to the increased flow energy of high Reynolds number, which is decomposing 2P modes shed close to the cylinder into a single P and S mode. The devolution of the 2P mode would have to be rapid as the sampling line of the dataset is located at a distance of  $4D$  in the wake of the cylinder. If the decomposition of the 2P mode is the mechanism by which the P+S mode appears in the dataset, it would have little effect on the global behaviour of the cylinder due to the observed rapid decay.

## 7.5 Summary

The chapter presents the application of the validated unsupervised clustering strategies to a case of high Reynolds number where the mapped domain is unknown to gain insights into the complex flow regimes. The quality of the clustering analysis was evaluated using internal clustering metrics, visual analysis of the clusters, and finally exploring the patterns of the generated vortex shedding maps.

The hybrid methods B and C were determined to provide the best results by outperforming single-stage methods in the quality of the clusters based on similarity of shape. The vortex shedding maps produced by the two hybrid methods provided valuable insights into the underlying dynamical regimes of the physical system. Specifically, the Reynolds case of 10,000 has similar vortex shedding modes as the low Reynolds number dataset, including the identified 2S and 2PO modes. The region of the map previously inhabited with 2P modes was observed to be comprised of the majority P+S modes. Overall, the flow physics derived from the cluster analysis demonstrates the increased dissipation effect of high Reynolds number flows, resulting in more variation and smoothing of the flow signals.

In conclusion, this chapter presented the use case of the data-driven vortex shedding map generating method using the reduced data source of local flow measurement of the  $y$ -component of velocity. The ability to extract meaningful clusters from more complex vortex structures that become increasingly indistinguishable was demonstrated using hybrid methods B and C. This data-driven method's versatility and performance yield exceptional results and significantly improve the vortex shedding map generation method due to reducing the data input and supervision required.

# Chapter 8

## Conclusions and Future Work

This thesis developed a data-driven approach for generating vortex shedding maps of a cylinder undergoing forced vibration. The wake classification strategy using machine learning (Chapter 5) demonstrated the ability to differentiate global vortex structures from local flow measurements in the wake of an oscillating cylinder. The unsupervised clustering of local flow measurement time series subsequences reproduced the benchmark vortex shedding map at Reynolds number 4000 (Chapter 6). The clustering method developed was then extrapolated to a high Reynolds number to produce a previously absent vortex shedding map (Chapter 7).

The presented wake classification strategy demonstrated the ability to identify global vortex structures using local flow signatures in Chapter 5. The proposed strategy demonstrated its accuracy in identifying vortex structures from a reduced input feature space. The improved feature space separation from using  $y$ -component of the velocity ( $u_y$ ) sensors result in the most improved testing accuracy ( $>15\%$ ) compared to the next best quantity, vorticity. The feature vector dataset derived from the  $y$ -component of the velocity ( $u_y$ ) sensors achieved testing accuracies of 99.3% and 99.8% using the random forest and  $k$ -nearest neighbour models, respectively. The four best performing machine learning models were selected for noise analysis which revealed that the random forest algorithm was the most robust to data corruption with a maximum reduction of 11.7% for the CvD case at the maximum noise level. Combining the results, the random forest classification algorithm (consisting of 107 estimators) was determined as the most advantageous machine learning model due to the balance of testing accuracy and reduced effect from noise. The strategy of wake classification provides a valuable tool that can be implemented in experimental setups to aid in controlling the behaviour of an oscillating cylinder. The methods were illustrated with the example of a bladeless wind turbine, but we believe that the methods apply more generally to any case of classification of vortex shedding. Additionally, the classification task acts as a vital proof of concept that the global vortex shedding modes can be deduced from the structure of the local flow measurements time series dataset. Furthermore, the use of the  $y$ -component of the velocity ( $u_y$ ) dataset will provide the best feature separation imperative for clustering time series data.

The procedure of generating vortex shedding maps using novel unsupervised subsequence clustering methods was presented and validated for the case of a low Reynolds number of 4000 in Chapter 6. Several clustering methods were selected and compared for the clustering task of subsequences extracted from the  $y$ -component of the velocity ( $u_y$ ) time series data. The published vortex shedding map by Morse and Williamson [7] at Reynolds number 4000 was used to validate the maps generated using the data-driven methods. The clustering results of the proposed methods demonstrated their ability to extract meaningful clusters that represented the underlying flow physics of the varying modes. The application of the clustering analysis to produce regime maps provided satisfactory agreement with the reference map by Morse and Williamson [7]. The proposed methods were then extended in Chapter 7 to quantify their performance to produce vortex shedding maps at high Reynolds numbers with more complex vortex structures. The vortex shedding maps produced at high Reynolds number provided novel insights on the underlying vortex structure interactions and identified numerous regions of similar patterns despite increased instability. The clustering procedure in the generation of vortex shedding maps offers a method



that requires less data and supervision without resolving the entire flow field. The method was implemented on the vortex shedding patterns of an oscillating cylinder, but the method could create value for any vortex shedding behaviour.

## 8.1 Future work

This thesis provides insight into how data-driven methods can be leveraged to identify and cluster vortex shedding signatures to produce vortex shedding maps, which are integral in the study of the VIV phenomenon. The results of the wake classification strategy could be expanded for more complex shedding modes or conducted at varying Reynolds numbers where the flow transitions into even more complex flow regimes. Furthermore, the noise analysis could be extended to higher levels of corruption or explore other sensor noise representations.

The next step in this stream of research is to extensively sample the normalized amplitude wavelength plane, specifically around the areas of transition zones, to gain a higher resolution of the cluster regions. There is also an opportunity to repeat the clustering procedure on numerical data obtained with large-eddy simulations (LES), which would provide higher resolution fluid measurement data. The improved resolution of the vortex structures is expected only to benefit the clustering and classification results. Finally, the frequency domain feature vector utilized in the wake classification method in Chapter 5 could be applied to the clustering technique to compare the clustering results from two varying methods.

# References

- [1] D. Fan, B. Wu, D. Bachina, and M. S. Triantafyllou, “Vortex-induced vibration of a piggyback pipeline half buried in the seabed,” *J. Sound Vib.*, vol. 449, pp. 182–195, Jun. 2019, doi: 10.1016/j.jsv.2019.02.038.
- [2] N. Kumar, V. Kumar Varma Kolahalam, M. Kantharaj, and S. Manda, “Suppression of vortex-induced vibrations using flexible shrouding—An experimental study,” *J. Fluids Struct.*, vol. 81, pp. 479–491, Aug. 2018, doi: 10.1016/j.jfluidstructs.2018.04.018.
- [3] W. Wang, X. Wang, X. Hua, G. Song, and Z. Chen, “Vibration control of vortex-induced vibrations of a bridge deck by a single-side pounding tuned mass damper,” *Eng. Struct.*, vol. 173, pp. 61–75, Oct. 2018, doi: 10.1016/j.engstruct.2018.06.099.
- [4] B. Jiang, X. Zhang, X. Xiao, and L. Zhang, “Vortex-induced vibration of a tube array with a large pitch-to-diameter ratio value,” *Adv. Mech. Eng.*, vol. 8, no. 7, p. 168781401665460, Jun. 2016, doi: 10.1177/1687814016654604.
- [5] T. Zhou, S. F. Mohd. Razali, Z. Hao, and L. Cheng, “On the study of vortex-induced vibration of a cylinder with helical strakes,” *J. Fluids Struct.*, vol. 27, no. 7, pp. 903–917, Oct. 2011, doi: 10.1016/j.jfluidstructs.2011.04.014.
- [6] MECA Enterprises Inc., *Helical-Strakes-windflowgraphic.jpg*. 2019. Accessed: Oct. 18, 2021. [Online]. Available: <https://www.mecaenterprises.com/wp-content/uploads/Helical-Strakes-windflowgraphic.jpg>
- [7] T. L. Morse and C. H. K. Williamson, “Fluid forcing, wake modes, and transitions for a cylinder undergoing controlled oscillations,” *J. Fluids Struct.*, vol. 25, no. 4, pp. 697–712, May 2009, doi: 10.1016/j.jfluidstructs.2008.12.003.
- [8] M. P. Païdoussis, S. J. Price, and E. de Langre, “3. Vortex-Induced Vibrations,” in *Fluid-Structure Interactions - Cross-Flow-Induced Instabilities*, Cambridge University Press. [Online]. Available: <https://app.knovel.com/hotlink/pdf/id:kt008NA8H1/fluid-structure-interactions/two-dimensional-viv-phenomenology>
- [9] Alexandra Techet, “13.42 Lecture: Vortex Induced Vibrations,” MIT OCW, Apr. 21, 2005. Accessed: Oct. 18, 2021. [Online]. Available: [https://ocw.mit.edu/courses/mechanical-engineering/2-22-design-principles-for-ocean-vehicles-13-42-spring-2005/readings/lec20\\_viv1.pdf](https://ocw.mit.edu/courses/mechanical-engineering/2-22-design-principles-for-ocean-vehicles-13-42-spring-2005/readings/lec20_viv1.pdf)
- [10] Michael S. Triantafyllou, Rémi Bourguet, Jason Dahl, and Yahya Modarres-Sadeghi, “Chapter 36. Vortex-Induced Vibrations,” in *Springer Handbook of Ocean Engineering*, 1st ed., Springer International Publishing, 2016, pp. 819–850.
- [11] W. Yang, E. Masroor, and M. A. Stremler, “The wake of a transversely oscillating circular cylinder in a flowing soap film at low Reynolds number,” *J. Fluids Struct.*, vol. 105, p. 103343, Aug. 2021, doi: 10.1016/j.jfluidstructs.2021.103343.
- [12] C. H. K. Williamson and R. Govardhan, “Vortex-induced vibrations,” *Annu. Rev. Fluid Mech.*, vol. 36, no. 1, pp. 413–455, 2004.
- [13] T. L. Morse and C. H. K. Williamson, “Employing controlled vibrations to predict fluid forces on a cylinder undergoing vortex-induced vibration,” *J. Fluids Struct.*, vol. 22, no. 6–7, pp. 877–884, Aug. 2006, doi: 10.1016/j.jfluidstructs.2006.04.004.

- [14]C. H. K. Williamson and A. Roshko, "Vortex formation in the wake of an oscillating cylinder," *J. Fluids Struct.*, vol. 2, no. 4, pp. 355–381, Jul. 1988, doi: 10.1016/S0889-9746(88)90058-8.
- [15]R. E. D. Bishop and A. Y. Hassan, "The Lift and Drag Forces on a Circular Cylinder Oscillating in a Flowing Fluid," *Proc. R. Soc. Lond. Ser. Math. Phys. Sci.*, vol. 277, no. 1368, pp. 51–75, 1964.
- [16]W. Wu, M. M. Bernitsas, and K. Maki, "RANS Simulation Versus Experiments of Flow Induced Motion of Circular Cylinder With Passive Turbulence Control at  $35,000 < RE < 130,000$ ," *J. Offshore Mech. Arct. Eng.*, vol. 136, no. 4, p. 041802, Nov. 2014, doi: 10.1115/1.4027895.
- [17]D. Zhang, W. Wang, H. Sun, and M. M. Bernitsas, "Influence of turbulence intensity on vortex pattern for a rigid cylinder with turbulence stimulation in flow induced oscillations," *Ocean Eng.*, vol. 237, p. 109349, Oct. 2021, doi: 10.1016/j.oceaneng.2021.109349.
- [18]S. Aghabozorgi, A. Seyed Shirkorshidi, and T. Ying Wah, "Time-series clustering – A decade review," *Inf. Syst.*, vol. 53, pp. 16–38, Oct. 2015, doi: 10.1016/j.is.2015.04.007.
- [19]L. Ye and E. Keogh, "Time series shapelets: a new primitive for data mining," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*, Paris, France, 2009, p. 947. doi: 10.1145/1557019.1557122.
- [20]G. Z. Hertz and G. D. Stormo, "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences," *Bioinformatics*, vol. 15, no. 7, pp. 563–577, Jul. 1999, doi: 10.1093/bioinformatics/15.7.563.
- [21]M. K. Das and H.-K. Dai, "A survey of DNA motif finding algorithms," *BMC Bioinformatics*, vol. 8, no. S7, p. S21, Dec. 2007, doi: 10.1186/1471-2105-8-S7-S21.
- [22]A. McGovern, D. H. Rosendahl, R. A. Brown, and K. K. Droegemeier, "Identifying predictive multi-dimensional time series motifs: an application to severe weather prediction," *Data Min. Knowl. Discov.*, vol. 22, no. 1–2, pp. 232–258, Jan. 2011, doi: 10.1007/s10618-010-0193-7.
- [23]Y. J. Fan and C. Kamath, "Identifying and Exploiting Diurnal Motifs in Wind Generation Time Series Data," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 29, no. 2, pp. 1550012-1-1550012–25, 2015.
- [24]L. Chi, Y. Feng, H. Chi, and Y. Huang, "Face image recognition based on time series motif discovery," in *2012 IEEE International Conference on Granular Computing*, Aug. 2012, pp. 72–77. doi: 10.1109/GrC.2012.6468574.
- [25]P. Beaudoin, S. Coros, M. van de Panne, and P. Poulin, "Motion-motif graphs," in *Proceedings of the 2008 ACM SIGGRAPH/Eurographics Symposium on computer animation*, 2008, pp. 117–126.
- [26]H. Sivaraks and C. A. Ratanamahatana, "Robust and Accurate Anomaly Detection in ECG Artifacts Using Time Series Motif Discovery," *Comput. Math. Methods Med.*, vol. 2015, pp. 1–20, 2015, doi: 10.1155/2015/453214.
- [27]E. Keogh and J. Lin, "Clustering of time-series subsequences is meaningless: implications for previous and future research," *Knowl. Inf. Syst.*, vol. 8, no. 2, pp. 154–177, 2005.
- [28]J. R. Chen, "Making clustering in delay-vector space meaningful," *Knowl. Inf. Syst.*, vol. 11, no. 3, pp. 369–385, Apr. 2007, doi: 10.1007/s10115-006-0042-6.
- [29]S. Torkamani and V. Lohweg, "Survey on time series motif discovery," *WIREs Data Min. Knowl. Discov.*, vol. 7, no. 2, p. e1199, 2017, doi: 10.1002/widm.1199.
- [30]A. Mueen, "Time series motif discovery: dimensions and applications," *WIREs Data Min. Knowl. Discov.*, vol. 4, no. 2, pp. 152–159, 2014, doi: 10.1002/widm.1119.
- [31]E. E. Özkoç, *Clustering of Time-Series Data*. IntechOpen, 2020. doi: 10.5772/intechopen.84490.
- [32]J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing SAX: a novel symbolic representation of time series," *Data Min. Knowl. Discov.*, vol. 15, no. 2, pp. 107–144, Aug. 2007, doi: 10.1007/s10618-007-0064-z.
- [33]Y.-L. Wu, D. Agrawal, and A. El Abbadi, "A comparison of DFT and DWT based similarity search in time-series databases," in *Proceedings of the ninth international conference on Information and knowledge management - CIKM '00*, McLean, Virginia, United States, 2000, pp. 488–495. doi: 10.1145/354756.354857.
- [34]Jiawei. Han, *Data mining concepts and techniques*, 3rd ed. Burlington, Mass: Elsevier, 2012.

- [35] S. Zolhavarieh, S. Aghabozorgi, and Y. W. Teh, “A Review of Subsequence Time Series Clustering,” *Sci. World J.*, vol. 2014, pp. 1–19, 2014, doi: 10.1155/2014/312521.
- [36] Romain Tavenard, *dtw\_vs\_euc.svg*. 2021. Accessed: Oct. 18, 2021. [Online]. Available: <https://rtavenar.github.io/blog/dtw.html>
- [37] L. Kaufman and P. J. Rousseeuw, Eds., “Partitioning Around Medoids (Program PAM),” in *Wiley Series in Probability and Statistics*, Hoboken, NJ, USA: John Wiley & Sons, Inc., 1990, pp. 68–125. doi: 10.1002/9780470316801.ch2.
- [38] P. Contreras and F. Murtagh, Eds., “Chapter 26: Hierarchical Clustering,” in *Handbook of Cluster Analysis*, New York: Chapman and Hall/CRC, 2015. doi: 10.1201/b19706.
- [39] M. Ester, H.-P. Kriegel, and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” in *AAAI Press*, Portland OR, 1996, pp. 226–231.
- [40] C. M. Bishop, *Pattern recognition and machine learning*. New York: Springer, 2006.
- [41] T. Kohonen, “Self-organized formation of topologically correct feature maps,” *Biol. Cybern.*, vol. 43, no. 1, pp. 59–69, 1982, doi: 10.1007/BF00337288.
- [42] C.-P. Lai, P.-C. Chung, and V. S. Tseng, “A novel two-level clustering method for time series data analysis,” *Expert Syst. Appl.*, vol. 37, no. 9, pp. 6319–6326, Sep. 2010, doi: 10.1016/j.eswa.2010.02.089.
- [43] X. Zhang, J. Liu, Y. Du, and T. Lv, “A novel clustering method on time series data,” *Expert Syst. Appl.*, vol. 38, no. 9, pp. 11891–11900, Sep. 2011, doi: 10.1016/j.eswa.2011.03.081.
- [44] S. Aghabozorgi, T. Ying Wah, T. Herawan, H. A. Jalab, M. A. Shaygan, and A. Jalali, “A Hybrid Algorithm for Clustering of Time Series Data Based on Affinity Search Technique,” *Sci. World J.*, vol. 2014, pp. 1–12, 2014, doi: 10.1155/2014/562194.
- [45] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, “Understanding of Internal Clustering Validation Measures,” in *2010 IEEE International Conference on Data Mining*, Sydney, Australia, Dec. 2010, pp. 911–916. doi: 10.1109/ICDM.2010.35.
- [46] C. Hennig, M. Meila, F. Murtagh, and R. Rocci, Eds., “Chapter 26: Method-Independent Indices for Cluster Validation and Estimating the Number of Clusters,” in *Handbook of Cluster Analysis*, New York: Chapman and Hall/CRC, 2015. doi: 10.1201/b19706.
- [47] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7.
- [48] J. C. Dunn†, “Well-Separated Clusters and Optimal Fuzzy Partitions,” *J. Cybern.*, vol. 4, no. 1, pp. 95–104, Jan. 1974, doi: 10.1080/01969727408546059.
- [49] David Yáñez Villareal, “VIV resonant wind generators,” Vortex Bladeless, Ávila, Spain, Jun. 2018. [Online]. Available: <https://vortexbladeless.com/download-green-paper/>
- [50] M. M. Bernitsas, K. Raghavan, Y. Ben-Simon, and E. M. H. Garcia, “VIVACE (Vortex Induced Vibration Aquatic Clean Energy): A New Concept in Generation of Clean and Renewable Energy From Fluid Flow,” *J. Offshore Mech. Arct. Eng.*, vol. 130, no. 4, pp. 41101–, 2008.
- [51] D. Mella, W. Brevis, and L. Susmel, “Spanwise wake development of a bottom-fixed cylinder subjected to vortex-induced vibrations,” *Ocean Eng.*, vol. 218, p. 108280, Dec. 2020, doi: 10.1016/j.oceaneng.2020.108280.
- [52] D. Lucor, J. Foo, and G. E. Karniadakis, “Vortex mode selection of a rigid cylinder subject to VIV at low mass-damping,” *J. Fluids Struct.*, vol. 20, no. 4, pp. 483–503, May 2005, doi: 10.1016/j.jfluidstructs.2005.02.002.
- [53] Z. Xiong and X. Liu, “Very Large-Eddy Simulations of the Flow Past an Oscillating Cylinder at a Subcritical Reynolds Number,” *Appl. Sci.*, vol. 10, no. 5, Art. no. 5, Jan. 2020, doi: 10.3390/app10051870.
- [54] E. Konstantinidis and D. Bouris, “Vortex synchronization in the cylinder wake due to harmonic and non-harmonic perturbations,” *J. Fluid Mech.*, vol. 804, pp. 248–277, Oct. 2016, doi: <http://dx.doi.org.proxy.lib.uwaterloo.ca/10.1017/jfm.2016.527>.
- [55] L. Deng, Y. Wang, Y. Liu, F. Wang, S. Li, and J. Liu, “A CNN-based vortex identification method,” *J. Vis.*, vol. 22, no. 1, pp. 65–78, Feb. 2019, doi: 10.1007/s12650-018-0523-1.

- [56]Z. Ye, Q. Chen, Y. Zhang, J. Zou, and Y. Zheng, “Identification of Vortex Structures in Flow Field Images Based on Convolutional Neural Network and Dynamic Mode Decomposition,” *Trait. Signal*, vol. 36, no. 6, pp. 501–506, Dec. 2019, doi: 10.18280/ts.360604.
- [57]X. Bai, C. Wang, and C. Li, “A Streampath-Based RCNN Approach to Ocean Eddy Detection,” *IEEE Access*, vol. 7, pp. 106336–106345, 2019, doi: 10.1109/ACCESS.2019.2931781.
- [58]B. Colvert, M. Als Salman, and E. Kanso, “Classifying vortex wakes using neural networks,” *Bioinspir. Biomim.*, vol. 13, no. 2, p. 025003, Feb. 2018, doi: 10.1088/1748-3190/aaa787.
- [59]M. Als Salman, B. Colvert, and E. Kanso, “Training bioinspired sensors to classify flows,” *Bioinspir. Biomim.*, vol. 14, no. 1, p. 016009, Nov. 2018, doi: 10.1088/1748-3190/aaef1d.
- [60]M. Wang and M. S. Hemati, “Detecting exotic wakes with hydrodynamic sensors,” *Theor. Comput. Fluid Dyn.*, vol. 33, no. 3–4, pp. 235–254, Aug. 2019, doi: 10.1007/s00162-019-00493-z.
- [61]F. J. Huera-Huarte and A. Vernet, “Vortex modes in the wake of an oscillating long flexible cylinder combining POD and fuzzy clustering,” *Exp. Fluids*, vol. 48, no. 6, pp. 999–1013, Jun. 2010, doi: 10.1007/s00348-009-0786-3.
- [62]K. Menon and R. Mittal, “Quantitative analysis of the kinematics and induced aerodynamic loading of individual vortices in vortex-dominated flows: A computation and data-driven approach,” *J. Comput. Phys.*, vol. 443, p. 110515, Oct. 2021, doi: 10.1016/j.jcp.2021.110515.
- [63]A. G. Calvet, M. Dave, and J. A. Franck, “Unsupervised clustering and performance prediction of vortex wakes from bio-inspired propulsors,” *Bioinspir. Biomim.*, vol. 16, no. 4, p. 046015, Jul. 2021, doi: 10.1088/1748-3190/ac011f.
- [64]M. Cann, R. McConkey, F.-S. Lien, W. Melek, and E. Yee, “Mode classification for vortex shedding from an oscillating wind turbine using machine learning,” *J. Phys. Conf. Ser.*, vol. 2141, no. 1, pp. 12009-, 2021.
- [65]Ryley, McConkey, “Vortex shedding in a turbulent wake.” Oct. 16, 2021. [Online]. Available: <https://www.kaggle.com/ryleymcconkey/vortex-shedding-wake>
- [66]J.-Z. Wu, H.-Y. Ma, and M.-D. Zhou, *Vorticity and vortex dynamics*. Berlin ; New York: Springer, 2006.
- [67]Ryley McConkey, *2S vortex shedding in the wake of an oscillating cylinder*, (Oct. 20, 2021). Accessed: Dec. 04, 2021. [Online Video]. Available: [https://www.youtube.com/watch?v=10sLGXmH\\_vQ](https://www.youtube.com/watch?v=10sLGXmH_vQ)
- [68]C.-C. M. Yeh *et al.*, “Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets,” in *2016 IEEE 16th International Conference on Data Mining (ICDM)*, Barcelona, Spain, Dec. 2016, pp. 1317–1322. doi: 10.1109/ICDM.2016.0179.
- [69]H. Li, Y. J. Wu, S. Zhang, and J. Zou, “Temporary rules of retail product sales time series based on the matrix profile,” *J. Retail. Consum. Serv.*, vol. 60, p. 102431, May 2021, doi: 10.1016/j.jretconser.2020.102431.
- [70]R. Wankhedkar and S. K. Jain, “Motif Discovery and Anomaly Detection in an ECG Using Matrix Profile,” in *Progress in Advanced Computing and Intelligent Engineering*, Singapore: Springer Singapore, 2020, pp. 88–95.
- [71]M. Zymbler and E. Ivanova, “Matrix Profile-Based Approach to Industrial Sensor Data Analysis Inside RDBMS,” *Math. Basel*, vol. 9, no. 17, pp. 2146-, 2021.
- [72]Abdullah Al Mueen and Earmonn Keogh, “Time Series Data Mining Using the Matrix Profile: A unifying View of Motif Discovery, Anomaly Detection, Segmentation, Classification, Clustering, and Similarity Joins,” *KDD 2017*, 2017.
- [73]F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *ArXiv12010490 Cs*, vol. 12, pp. 2825–2830, Jun. 2018.
- [74]D. Arthur and S. Vassilvitskii, “k-means++: The Advantages of Careful Seeding,” p. 11.
- [75]S. Aghabozorgi and Y. W. Teh, “Stock market co-movement assessment using a three-phase clustering method,” *Expert Syst. Appl.*, vol. 41, no. 4, Part 1, pp. 1301–1314, Mar. 2014, doi: 10.1016/j.eswa.2013.08.028.

- [76]K. W. Lin and C.-H. Lin, “A fast CAST-based clustering algorithm for very large database,” in *Proceedings 2011 International Conference on System Science and Engineering*, Jun. 2011, pp. 420–424. doi: 10.1109/ICSSE.2011.5961940.
- [77]S. L. Brunton, M. S. Hemati, and K. Taira, “Special issue on machine learning and data-driven methods in fluid dynamics,” *Theor. Comput. Fluid Dyn.*, vol. 34, no. 4, pp. 333–337, Aug. 2020, doi: 10.1007/s00162-020-00542-y.
- [78]M. Arslan, M. Güzel, M. Demirci, and S. Ozdemir, “SMOTE and Gaussian Noise Based Sensor Data Augmentation,” 2019. doi: 10.1109/UBMK.2019.8907003.
- [79]L. J. P. van der Maaten and G. E. Hinton, “Visualizing High-Dimensional Data Using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, no. nov, pp. 2579–2605, 2008.
- [80]A. L. N. Fred and J. M. N. Leitão, “Partitional vs Hierarchical Clustering Using a Minimum Grammar Complexity Approach,” in *Advances in Pattern Recognition*, vol. 1876, F. J. Ferri, J. M. Iñesta, A. Amin, and P. Pudil, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 193–202. doi: 10.1007/3-540-44522-6\_20.
- [81]Douglas Steinley, “Chapter 4: K-Medoids and Other Criteria for Crisp Clustering,” in *Handbook of Cluster Analysis*, New York: Chapman and Hall/CRC, 2015.