

Dissertation

Statistical and Fractal Analysis of Particle Data
from Two-Dimensional Video Disdrometer

Graduate School of
Natural Science & Technology
Kanazawa University

Division of Electrical Engineering
and Computer Science

Student ID No.: 1223112004

Name: Gavrilov Sergey

Chief advisor: Professor Kenji Satou

Date of Submission: January, 2015

Abstract

Precipitation (or its absence) plays a major role in almost any ecosystem as water balance and circulation are crucial factors for most living biological species. Studies concerning precipitation are common in the field of meteorology as contributing to a range of purposes in such areas as agriculture, disaster prevention and others.

Depending on the location, the intensity and types of precipitation greatly varies. In countries having a cold winter season with rather low temperature or areas of high latitude, the studies of solid precipitation are being conducted mainly in order to decrease the potential damage caused by heavy snowfall.

Recently numerous advances in meteorological field and weather forecasting had been achieved due to the development of high-end optics, electronics, and raise of computational power. However, important issues such as the snowfall formation mechanism and accurate winter precipitation quantifying are still lacking deeper studies.

In this study, we intended to develop a novel system for particle-by-particle observation and accurate classification of solid precipitation particles explicitly using their shape and performing fractal analysis. This thesis aims to enhance the accuracy in the microphysical parameterizations in numerical forecast models.

In order to conduct field experiments, a ground observation system for solid precipitation using two-dimensional video disdrometer (2DVD) was developed on the roof of Kanazawa University. Among 16,010 particles observed by the system, around 10% of them were randomly sampled and manually classified into five classes, that is snowflake, snowflake-like, intermediate, graupel-like, and graupel. At first, each particle was represented by a vector of 72 features containing fractal dimension and box-count to represent the complexity of particle shape. Feature analysis on the dataset clarified the importance of fractal dimension and box-count features for characterizing particles varying from snowflakes to graupel. On the other hand, performance evaluation of two-class classification by Support Vector Machine (SVM) was conducted.

The experimental results revealed that, by selecting only 10 features out of 72, the average accuracy of classifying particles into snowflakes and graupel can reach

around 95.4%, which has not been achieved by previous studies. Although the 2DVD takes binary image with lower resolution than CCD video camera, combination of up-to-date classifier and features including fractal-related ones enabled the system to outperform the accuracy achieved in our previous study.

We believe that the approach described in this thesis may be adapted and implemented for the purpose of further research on understanding the snowfall formation processes contributing to reduce heavy snowfall damage.

Acknowledgments

Looking back at the accomplished work, I realize that it would not be done without the help and support of a large number of people. I am thankful to everyone who helped and encouraged me all along these three years. I hereby take the opportunity to express my gratitude and appreciate everyone who contributed to this work.

First, I would like to emphasize my great respect and deep gratitude to my supervisor, Professor Kenji Satou for all his enormous help and exceptional guidance throughout the research. He is the supervisor one can only wish for: motivating, encouraging and kind.

I owe my sincere gratitude to Professor Mamoru Kubo as to the person who taught me everything I know about meteorology and snow. His help goes far beyond the limits of an advisor, as he is the one who was always there to help and advice on any matter.

I feel proud and lucky to be able to study and work with these people as my advisors.

My gratitude also goes to all the staff of Kanazawa University, to everyone teaching me and cooperating with me.

Special thanks to President Ken-ichiro Muramoto of Ishikawa National College of Technology and Professor Yasushi Fujiyoshi of Hokkaido University for cooperation in the snowfall observation.

Furthermore, I am very thankful to Mitani Sangyo Eikai for the provided scholarship.

I am also very grateful to Doctor Ryoko Shibata for providing financial support that gave me the opportunity to live and study in Japan.

I would like to thank all of my surrounding: the students of Bioinformatics Laboratory, my Japanese friends as they provided me a warm atmosphere for living.

Last but not least, I would like to thank my family, especially my mother and my wife as those two are always cheering up and encouraging me.

Contents

| | |
|---|------|
| Abstract | i |
| Acknowledgments..... | iii |
| Contents | iv |
| List of Figures | vi |
| List of Tables | viii |
| Abbreviations | ix |
| Chapter 1 Introduction | 1 |
| 1.1 Research context..... | 1 |
| 1.1.1 Meteorology and weather monitoring | 1 |
| 1.1.2 Two-dimensional video disdrometer | 2 |
| 1.1.3 Types of solid precipitation | 4 |
| 1.2 Objectives | 5 |
| 1.3 Contributions | 5 |
| 1.4 Thesis organization..... | 7 |
| Chapter 2 Related Works and Objectives | 8 |
| 2.1 Snow classification methods overview | 8 |
| 2.2 Thesis organization | 10 |
| 2.3 Objectives | 12 |
| Chapter 3 Materials and Methods | 14 |
| 3.1 System and condition of observation | 14 |
| 3.2 Preparation of data for analysis and classification | 14 |
| 3.2.1 Particle images and basic features | 14 |
| 3.2.2 Fractal-related features | 20 |
| 3.2.3 Human annotation | 22 |
| 3.3 Algorithms | 24 |

| | | |
|--------------|---|----|
| 3.3.1 | Normalization | 24 |
| 3.3.2 | Pearson's correlation coefficient | 24 |
| 3.3.3 | Principal component analysis (PCA)..... | 24 |
| 3.3.4 | Support vector machine (SVM)..... | 25 |
| 3.3.5 | Cross-validation..... | 25 |
| Chapter 4 | Experimental Results and Discussion | 26 |
| 4.1 | Feature analysis by Pearson's correlation coefficient | 26 |
| 4.2 | Feature analysis by PCA | 28 |
| 4.3 | Particle classification by SVM | 42 |
| 4.4 | Classification of unlabeled data..... | 45 |
| Chapter 5 | Conclusion and Future Works | 58 |
| 5.1 | Dissertation summary | 58 |
| 5.2 | Future works | 59 |
| Bibliography | | 60 |

List of Figures

| | | |
|-------------|---|----|
| Figure 1.1 | 2DVD units organization..... | 3 |
| Figure 1.2 | A 2DVD Sensor unit architecture..... | 4 |
| Figure 1.3 | Example of snowflake image from 2DVD..... | 4 |
| Figure 2.1 | Overview of CCD video camera system used in previous research..... | 9 |
| Figure 2.2 | Overview of data processing flow..... | 10 |
| Figure 3.1 | Photograph of 2DVD sensor unit covered with snow..... | 15 |
| Figure 3.2 | MTSAT-2 satellite image at 1200 JST 26 January 2011..... | 16 |
| Figure 3.3 | Particle images taken by 2DVD..... | 17 |
| Figure 3.4 | Integration of camera-specific features into max and min values..... | 19 |
| Figure 3.5 | Example of covering results from the box-counting method..... | 21 |
| Figure 3.6 | The log-log plot of the box-counting method..... | 22 |
| Figure 4.1 | Correlation analysis of features..... | 27 |
| Figure 4.2 | PC1 of the datasets except “warning-only”..... | 29 |
| Figure 4.3 | PC2 of the datasets except “warning-only”..... | 29 |
| Figure 4.4 | PC3 of the datasets except “warning-only”..... | 30 |
| Figure 4.5 | 3D plots of PC1, PC2, and PC3 in the dataset “all” from three different angles of view..... | 35 |
| Figure 4.6 | 3D plots of PC1, PC2, and PC3 in the dataset “no-warning” from three different angles of view..... | 36 |
| Figure 4.7 | 3D plots of PC1, PC2, and PC3 in the dataset “5-classes” from three different angles of view..... | 37 |
| Figure 4.8 | 3D plots of PC1, PC2, and PC3 in the dataset “2-classes” from three different angles of view..... | 38 |
| Figure 4.9 | 3D plots of PC1, PC2, and PC3 in the dataset “warning” from three different angles of view..... | 39 |
| Figure 4.10 | 3D plots of the dataset “5-classes” without some features. In left and right panels, camera-specific and camera-independent box-count features are removed, respectively..... | 40 |

| | |
|--|----|
| Figure 4.11 3D plots of the dataset “5-classes” without some features. In left and right panels, camera-specific and camera-independent fractal features are removed, respectively. | 41 |
| Figure 4.12 3D plots of PC1, PC2, and PC3 in the dataset “unlabeled” from three different angles of view..... | 46 |
| Figure 4.13 3D plots of PC1, PC2, and time in the dataset “unlabeled” with the view from PC1-time plane. | 47 |
| Figure 4.14 3D plots of PC1, PC2, and time in the dataset “unlabeled” with the view from PC2-time plane. | 48 |
| Figure 4.15 3D plots of PC1, PC2, and time in the dataset “unlabeled” with the view from PC1-time plane. | 49 |
| Figure 4.16 3D plots of PC1, PC3, and time in the dataset “unlabeled” with the view from PC1-time plane. | 50 |
| Figure 4.17 3D plots of PC1, PC3, and time in the dataset “unlabeled” with the view from PC3-time plane. | 51 |
| Figure 4.18 3D plots of PC1, PC2, and time in the dataset “unlabeled” with the view from PC1-time plane. | 52 |
| Figure 4.19 3D plots of PC2, PC3, and time in the dataset “unlabeled” with the view from PC2-time plane. | 53 |
| Figure 4.20 3D plots of PC2, PC3, and time in the dataset “unlabeled” with the view from PC3-time plane. | 54 |
| Figure 4.21 3D plots of PC1, PC2, and time in the dataset “unlabeled” with the view from PC1-time plane. | 55 |
| Figure 4.22 Histogram of predicted snowflakes in “unlabeled” dataset..... | 56 |
| Figure 4.23 Histogram of predicted graupels in “unlabeled” dataset. | 57 |

List of Tables

| | | |
|-----------|--|----|
| Table 3.1 | Features for Analysis and Classification..... | 18 |
| Table 3.2 | The number of samples after annotation..... | 23 |
| Table 3.3 | Datasets according to annotation. | 23 |
| Table 4.1 | Top 10 features in descending order of PC1 values..... | 31 |
| Table 4.2 | Top 10 features in descending order of PC2 values..... | 32 |
| Table 4.3 | Top 10 features in descending order of PC3 values..... | 33 |
| Table 4.4 | Average errors in the predictions by single feature and multiple features with backward elimination..... | 44 |

Abbreviations

2DVD = Two-dimensional Video Disdrometer

CCD = Charge-Coupled Device

PCA = Principal Component Analysis

SVM = Support Vector Machine

HSE = Heavy Snowfall Events

QPF = Quantitative Precipitation Forecasting

MPS = Microphysical Parameterization Schemes

Chapter 1

Introduction

In this chapter we first introduce some basic aspects of meteorology concerning solid precipitation (snowfall). After that, we present current weather monitoring and prediction problems such as snow type identification and precipitation quantification. The purpose of this thesis is to deal with these problems. In the end of this chapter thesis contributions and further structure organization are provided.

1.1 Research context

1.1.1 Meteorology and weather monitoring

Due to the diversity of terrain, rainfall and snowfall phenomena take on different forms depending on location. The amount and type of precipitation may change quite rapidly over a short period of time[1][2].

As heavy snowfall may cause severe damage, it is a significant issue to be able to monitor precipitation continuously for decreasing the potential damage as well as obtaining a better meteorological understanding of orographic snowfall. Especially, it is important to understand the snowfall formation mechanism with different types of solid precipitation such as snowflake and graupel.

Modern meteorological weather monitoring consists of a large variety of approaches and techniques, using both remote (radars, lidars) and ground-based observation equipment and methods.

For the purpose of remote measuring the precipitation intensity on a wide area, a popular facility is a *polarimetric radar*[3][4][5][6]. This device is commonly

used to obtain the cloud microphysical parameters. While polarimetric radars operate on large-scale, a device named *disdrometer* is additionally used for the ground-based observation of precipitation at a spot. It is a relatively-small instrument which can measure the size and falling velocity of a particle. Based on the fact that rain and graupel have different distribution of size and falling velocity, it is possible to discriminate them using a disdrometer[7][8][9]. However, if two particles have similar size and falling velocity, it is impossible to discriminate them by a disdrometer. In this sense, the observation of precipitation using a polarimetric radar and/or a disdrometer is not sufficient for accurately estimating the amount of precipitation consisting of various types[10][11][12][13].

1.1.2 Two-dimensional video disdrometer

A two-dimensional video disdrometer (hereafter 2DVD) is an optical device developed for measuring solid precipitation characteristics on ground. The instrument is manufactured by Joanneum Research of Austria. 2DVD measures volume, diameter, shape, and velocity of every individual particle. From this data, one can estimate particle size distribution, precipitation rate, and other related variables[14][15].

The device consists of 3 units: a sensor unit hosting cameras and optics, an outdoor electronics unit that operates the cameras and sends data to the indoor processing unit. The equipment arrangement is shown on Figure 1.1.

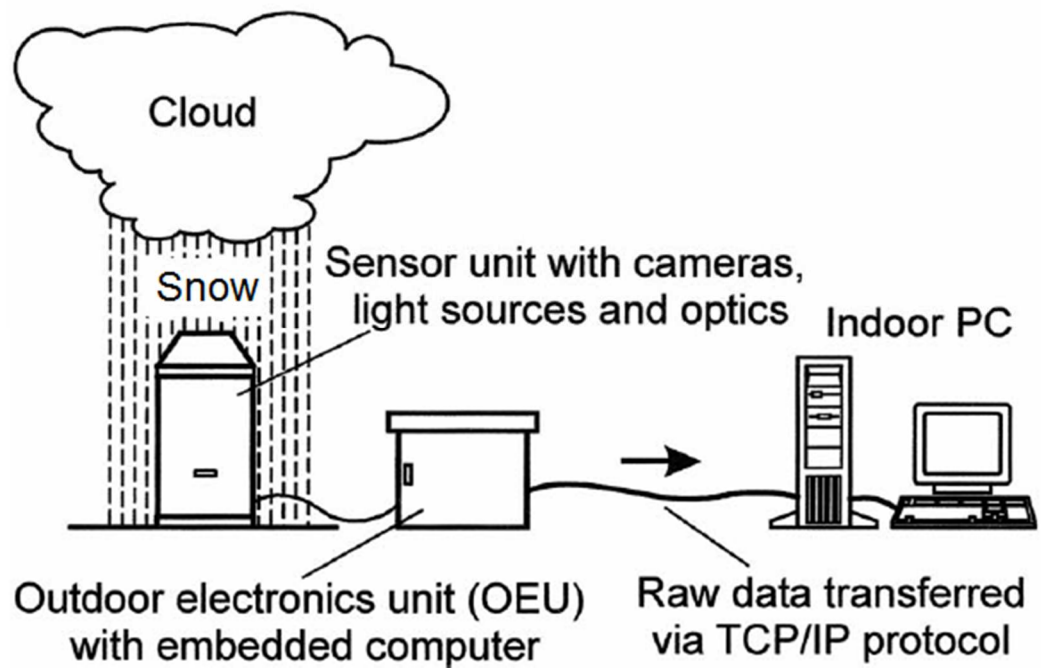


Figure 1.1 2DVD units organization (Original figure from [16]).

Figure 1.2 shows the 2DVD sensor unit inner structure. The sensor unit consists of two orthogonal and synchronized line-scan cameras and a bright light source in front of each of them. While precipitation particles fall between the cameras and light sources (an area of $10\text{cm} \times 10\text{cm}$), their shapes are recorded as shadows are being projected. Although they are low-resolution images of black and white (Figure 1.3), the obtained shape information is sufficient for particle classification.

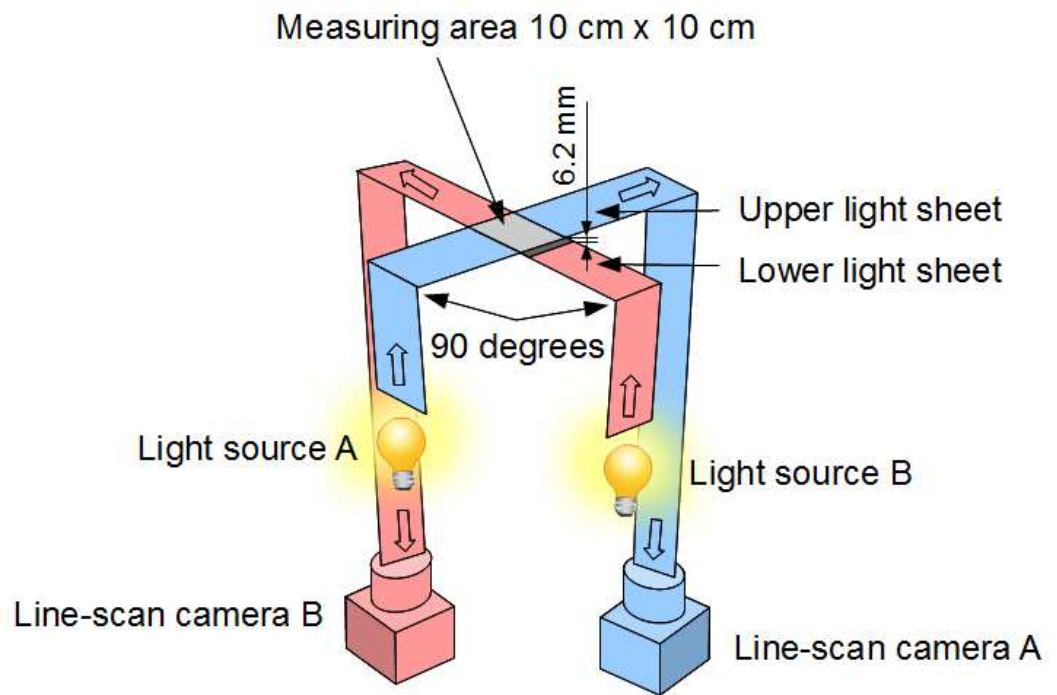


Figure 1.2 A 2DVD Sensor unit architecture.



Figure 1.3 Example of snowflake image from 2DVD.

1.1.3 Types of solid precipitation

While liquid precipitation consists of raindrops only, solid precipitation may be split into a variety of classes, depending on the particle parameters. These

parameters are influenced by various factors such as snow formation processes and macro physical conditions.

This study is intended to make difference only between hydrometeors of 2 basic classes: *snow* and *graupel*. Nevertheless it makes use of 3 intermediate classes which are artificial in the sense that are derived by manual annotation in difficult to classify cases.

A graupel is round-shaped as an approximate ellipse, and in contrast, a snowflake has a complex shape. As to the size of a particle, graupels are relatively smaller than snowflakes. These features meet intuitive criteria in human's discrimination of snowflake and graupel. The latter feature was frequently used in previous studies since it is easier to observe[16].

1.2 Objectives

As the classification of solid precipitation into snowflake and graupel is important for weather monitoring purposes, this thesis aims to solve the problem of improving the accuracy of that kind of studies implementing a new approach to enhance the results.

The main objective of this work is to create a novel method of particle-to-particle classification of solid precipitation into snowflake and graupel based on enriched information about the fractal properties of every single particle using 2DVD as primary data acquiring system and making explicit use of obtained particles shape features.

1.3 Contributions

Studies related to solid precipitation monitoring may contribute to a wide range of theoretical and practical issues, for instance snowfall formation mechanism or winter precipitation estimation to predict intensive snowfall.

This study may contribute to the following matters:

Decreasing the damage of heavy snowfall events (HSE).

Unexpected heavy snowfall may cause severe damage to large areas of crops, disrupt or block the traffic on roads and stop the indispensable facilities operation.[17] Furthermore, heavy snowfall may result in large floods after the snow melts.[18] Natural hazards are usually inevitable, so it makes extremely important the need to be able to forecast them as soon as possible to get prepared and minimize the potential risks and decrease the caused damage.

Understanding the snowfall formation mechanism.

Snowfall formation mechanism is a matter to which in recent years there has been marked scientific interest as foretelling the possible HSE is complicated due to the variety of impact factors such as elevation levels, climate change and so on. [19] Nevertheless we believe that the method proposed in this thesis is adaptive to the location specifications and various environmental conditions and may contribute to get a better understanding of the snowfall formation mechanism via additional acquired data.

Improvement microphysical parameterizations in numerical forecast models.

Despite of the recent significant improvements in numerical model resolution and major achievements in various meteorological parameters forecasting, the progress in the area of quantitative precipitation forecasting (QPF) has been relatively slow. That is also related to the lacking in microphysical parameterization schemes (MPS) used to make cloud models and precipitation processes simulations. In order to solve this problem, one needs to differentiate errors in QPF from MPS to those from other sources.[20] That data obtained from our study can be used to enhance the microphysical processes and hydrometeor fields simulated by MPS.

Quantifying winter precipitation accurately.

The need to increase solid precipitation measurement accuracy in all weather conditions remains challenging despite current state-of-the-art methodologies improvement. This is mostly due to high error in data acquisition during windy conditions (not all particles are being caught by measurement devices).[21]

Development algorithms for determining particle size distributions with remote sensors such as polarimetric radar.

Solid precipitation type distributions derived from a 2 dimensional video disdrometer might be used to find mutual relationships between particle size distribution parameters and to verify the polarimetric radar accuracy in hydrometeor type identification.[3] This makes ground based observations a useful tool for improving and checking the data acquired from radars. In this sense, the development of accurate and computationally easy algorithms is a challenging issue.

1.4 Thesis organization

The thesis consists of 5 chapters, including the current introduction chapter covering the background concepts, introducing the main research objectives, contributions and organization.

The remaining chapters are as follows:

Chapter 2 reviews state-of-the-art snow classification methods that contribute to nearly same objectives. Some of the main problems of research in this field are being addressed.

Chapter 3 introduces the proposed novel approach using fractal related features to enhance the accuracy of solid precipitation classification. Main methods and algorithms used in the study are also described in this chapter.

Chapter 4 shows and discusses the obtained experimental results and

Chapter 5 summarizes the thesis giving conclusion of achievements and discusses the future works vector.

Chapter 2

Related Works and Objectives

In this chapter, the review of the state-of-the-art snow classification methods is presented. It gives an overview of the previous works along comparison with the current thesis. It is followed by the common problems in the research field of solid precipitation classification and objective discussion.

2.1 Snow classification methods overview

Modern meteorological studies make use of the top-level devices and approaches not only to quantitatively evaluate the precipitation amount but to gain additional information about precipitation particles. This data may then be used to interpret images obtained from radars and enhance the understanding of particle microphysics, cloud formation processes and so on.

Liquid precipitation consists of raindrops only and can be easily distinguished from solid precipitation. In case of solid precipitation, it may be split into a variety of classes, depending on the particle parameters. The task to classify solid precipitation particles is still being an open problem and widely discussed in communities related to weather monitoring in countries with cold winter season.

Among the variety of works and studies related to this issue, we consider the following 2 papers to be the closest and most recent works in the field of snow classification:

1. Nurzynska, K., Kubo, M. and Muramoto, K. (2010) 2D Feature Space for Snow Particle Classification into Snowflake and Graupel. *IEICE Transactions on Information and Systems*, E93-D, **12**, 3344-3351. [16]

2. Grazioli, J., Tuia, D., Monhart, S., Schneebeli, M., Raupach, T. and Berne, A. (2014) Hydrometeor classification from two-dimensional video disdrometer data. *Atmospheric Measurement Techniques*, **7**, 2869-2882. [22]

First one is the previous work of the Bioinformatics Laboratory of Kanazawa University. In that study a deep analysis of feature importance, classifier differences and data flow impact had been conducted. Major attention is dedicated to the feature selection and gaining highest accuracy possible using pairs of features and different classifiers and data flow organization. The used optical device was a CCD video camera, which had been taking 1280×960 grayscale images. Using rich information of these high-resolution grayscale images, it achieved high accuracy (over 90%) of particle-by-particle classification into snowflake and graupel.

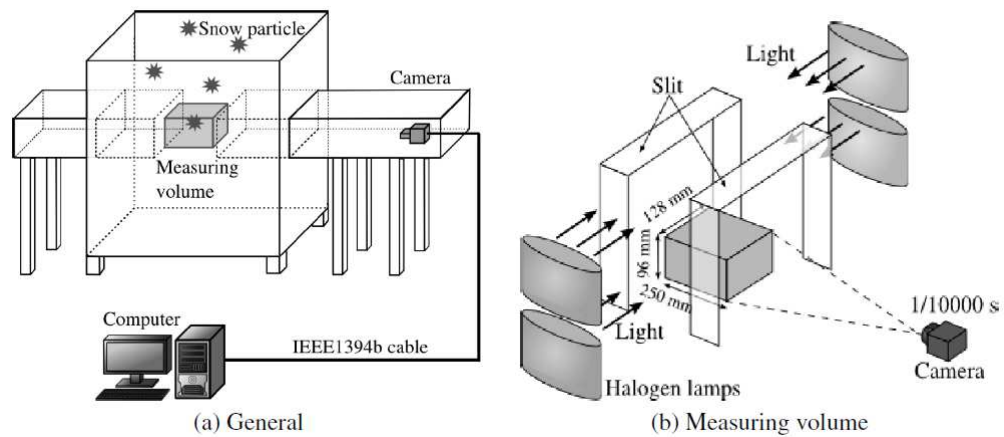


Figure 2.1 Overview of CCD video camera system used in previous research.
(From original paper by Nurzynska, K., Kubo, M. and Muramoto, K.)

Among the weak points one can mention that since the equipment set requires large space comparable by size to a small room ($2\text{m} \times 4\text{m}$), portability and applicability are rather low. In addition, it is a hand-made facility and not easy to adjust and use.

Comparing to that work the current study uses different equipment - 2DVD instead of a CCD video camera - and makes use of different features; explicitly using the fractal related ones. While on the other hand some features such as

brightness and Hu moment of the 1st order are unavailable due to the black and white source image.

Moreover, the classifiers selected to be used in the previous study were Mahalanobis Minimum Distance and k-Nearest Neighborhood. As classifiers algorithms used in the two studies differ it makes hard to estimate and compare the overall method results accuracy.

Our study aims to get a higher accuracy rate based on a set of features gained from low resolution black and white imagery.

In the second paper the authors used 2DVD to determine the dominant type of precipitation observed in a time interval. While using same equipment, the feature sets of the studies are different as the aims of the former are to select the main precipitation type rather than to get a full overview of the precipitation process. Conversely saying, it does not perform particle-by-particle classification which is the main basis of our study.

2.2 Thesis organization

Problems that arise during the whole research process are being described in order of this thesis process workflow (Figure 2.2) and are related to each particular step or method.

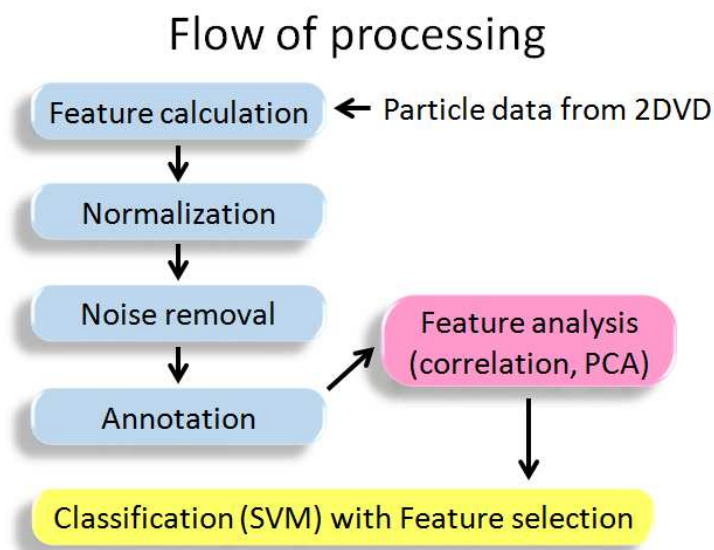


Figure 2.2 Overview of data processing flow.

Briefly, one can consider two main groups of problem – those arising during the data acquisition and the problems concerning data processing. Both groups are derived from limitations of hardware and methodology respectively.

Illumination

In order to obtain a high resolution precipitation particle image all the measuring volume should be properly enlightened. Good image-plane illuminance becomes critical as the exposure time may not be increased (shutter speed should be fast in order to operate online) and the tradeoff between these two variables to obtain needed photometric exposure is impossible.

The need of a good illumination system providing sufficient light was taken into account in both researches according to the used optical devices. In this sense line-scan cameras are more compact than CCD video cameras though provide less detailed information about particle shape.

Dealing with strong wind

One of the most challenging problems is to deal with the strong wind during the experiment. Box-shaped shields protecting the optical devices cause problems to snowfall particles evenly reach the entrance slit or hole to the measuring volume in the case of strong horizontal shift during precipitation.

The construction characteristics and included software of 2DVD are decreasing the influence of wind to the acquiring data still cannot deal with wind after a certain velocity threshold. This problem seems to be hard to deal with.

Manual annotation

Solid precipitation takes different forms depending geographic location and many other factors influencing snow formation mechanisms. Classification into snow and graupel adopted in this study and the previous work by Nurzynska et al. is appropriate for Kanazawa city (36.544°N, 136.705°E). In case of the work of Grazioli et al. the discrimination of 8 hydrometeor classes was used (small-particle-like, dendrite-like, column-like, graupel-like, rimed particle-like, aggregate-like, melting-snow-like, and rain).

Manual annotation requires an utter knowledge of solid precipitation types not only theoretical but applied to the location of interest as the production of a reliable training set significantly decreases the accuracy errors in further classification process. Such specialists are a rarity and our research team is lucky to have one in the person of Professor Mamoru Kubo.

Feature selection

One of the most challenging problems concerning the improvement in accuracy is which features to use. As it is seen from the results evaluation a large variety of shape related features do not provide efficient information for the classification purpose. Vice versa, some features that are not intuitively apparent may drastically contribute to it. That makes essential the need of finding new features and checking the existing.

The work by Nurzynska et al. focuses on pair wise feature analysis explicitly checking features classification value. In the study by Grazioli et al. most features are also those shape related which are used in remote sensing and new feature derivation for accuracy maximizing was not the ultimate goal [22]. In our work we are making use of the fractal properties and trying to estimate their contribution.

Classifier selection

Currently existing classifier algorithms are rather numerous and may show different results depending on the datasets. The investigation of the best of them for the use on the snowfall dataset is nontrivial and might be a separate topic. Nurzynska et al. study showed performance difference based on the different classifiers. Our research and the study by Grazioli et al. are using SVM due to its popularity, applicability and high-performance. The search of optimal classifier was not included in these research outlines and remains a case study.

2.3 Objectives

Speaking of the state-of-the-art projects and research studies related to solid precipitation classification one can see that their main objectives vary mostly depending on the area of practical contribution as this research area is tightly

connected to further real-world implementation enhancing current meteorological systems.

Dealing with these common problems that were mentioned in the previous subsection is subject of this kind of studies and this particular one. Achieving success in solving these complex tasks along with individual additional ones, leads to the proper solution of the prescribed objective.

The evaluation of the two works results shows that the particle-by-particle classification using the fractal-related shape describing features allows enhancing the classification accuracy at least by 4.5% confirming the objective of this thesis.

Chapter 3

Materials and Methods

In this chapter, various information about observation, data, algorithm, etc. used in this study is presented.

3.1 System and condition of observation

We have observed snowfall event from 1250 JST to 1300 JST in January 26, 2011 at Kanazawa University. The data of 16,010 snow particles were recorded by the 2DVD (Figure 3.1). Figure 3.2 shows MTSAT-2 satellite image at 1200 JST 26 January 2011 and the location of observation point. The air temperature was about 0°C through the event duration.

3.2 Preparation of data for analysis and classification

3.2.1 Particle images and basic features

Figure 3.3 illustrates examples of particle image data recognized and generated by 2DVD. Since 2DVD scans two line images at once from two orthogonally oriented cameras (A and B), two different images are obtained for each particle.

In Figure 3.3, it can be seen that a graupel is round-shaped as an approximate ellipse, and in contrast, a snowflake has a complex shape. As to the size of a particle, graupels are relatively smaller than snowflakes. These features meet intuitive criteria in human's discrimination of snowflake and graupel. The latter feature was frequently used in previous studies since it is easier to observe.

In addition to shape and size, it is possible to obtain various features of a particle by using 2DVD. The list of features used in this study is shown in Table 3.1.



Figure 3.1 Photograph of 2DVD sensor unit covered with snow.

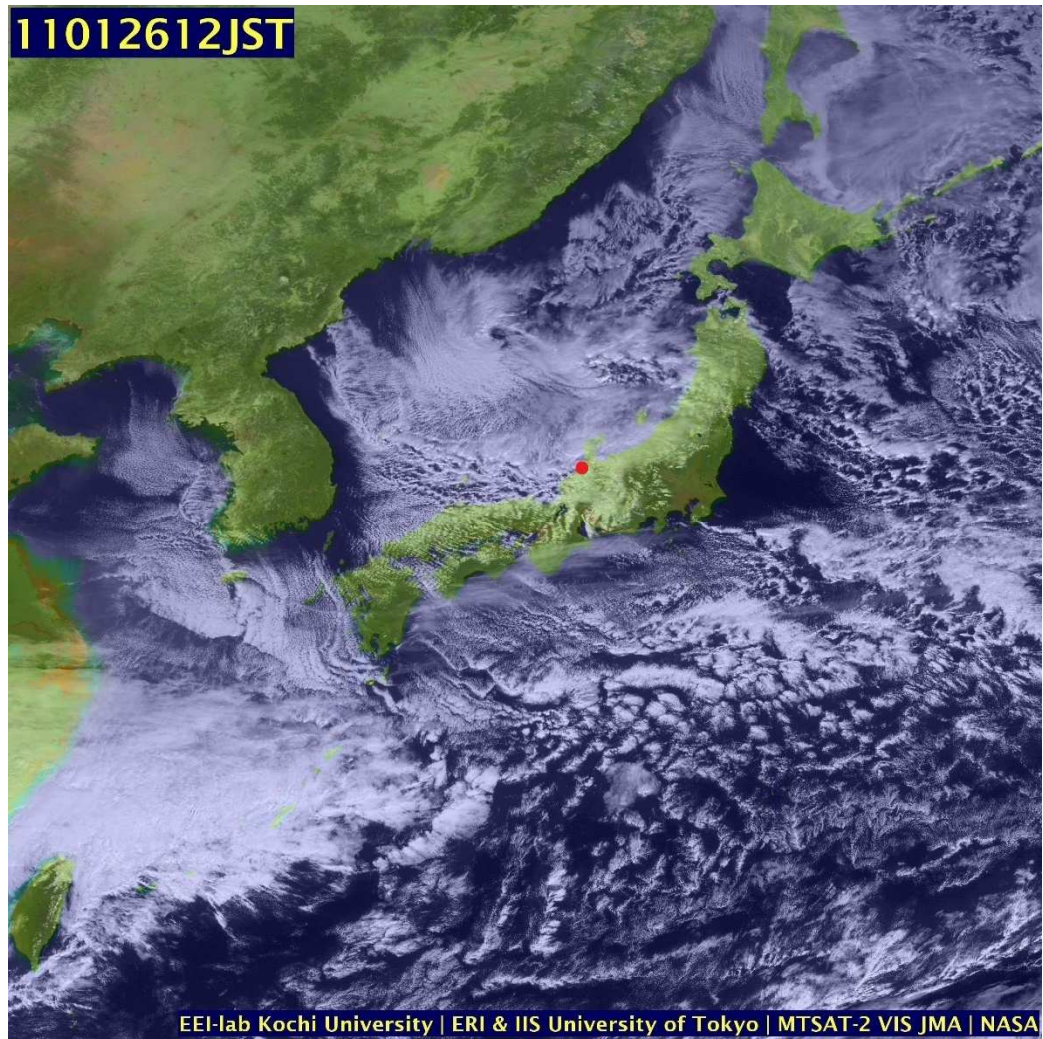


Figure 3.2 MTSAT-2 satellite image at 1200 JST 26 January 2011 (from <http://weather.is.kochi-u.ac.jp/>). The 2DVD is installed at Kanazawa University and the location of observation point is indicated by a red circle. (36.544°N, 136.705°E).

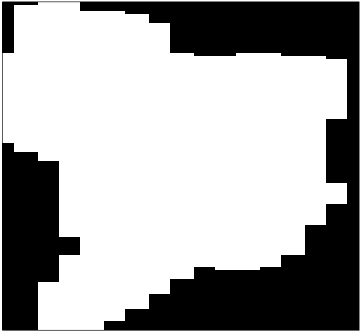
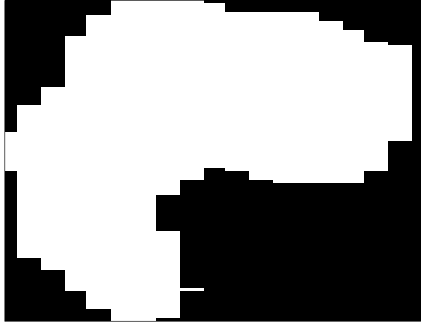
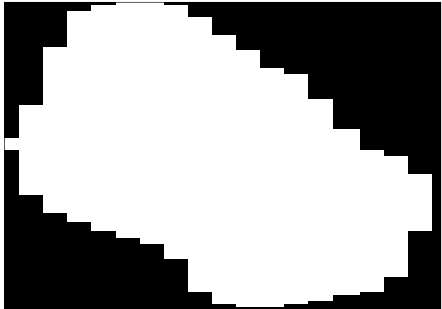
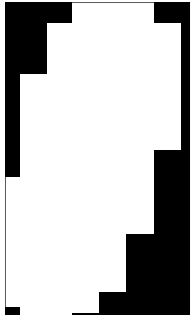
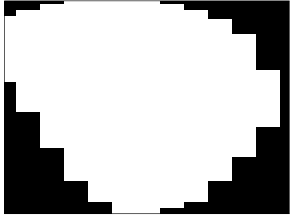
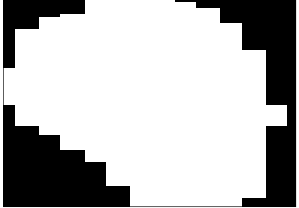
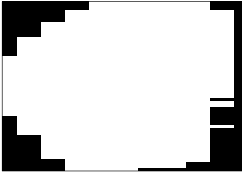
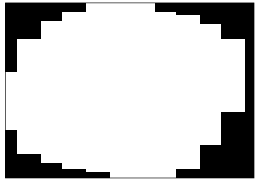
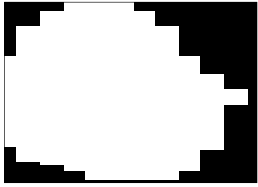
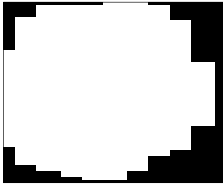
| | image by camera A | image by camera B |
|----------------|---|---|
| snowflake |  |  |
| snowflake-like |  |  |
| intermediate |  |  |
| graupel-like |  |  |
| graupel |  |  |

Figure 3.3 Particle images taken by 2DVD.

Table3.1 Features for Analysis and Classification.

| Feature type | Feature name |
|---|--|
| Camera-independent features | equivolumetric_diameter[mm], volume[mm ³], vertical_fall_velocity[m/s], height_of_one_line[mm] |
| Camera-specific features | height[mm]_A, height[mm]_B, number_of_lines_A, number_of_lines_B, pixelwidth[mm]_A, pixelwidth[mm]_B, width[pixel]_A, width[pixel]_B, height[pixel]_A, height[pixel]_B, total_pixels_A, total_pixels_B, area[mm ²]_A, area[mm ²]_B, perimeter[mm]_A, perimeter[mm]_B, box_count_1_A, box_count_1_B, box_count_2_A, box_count_2_B, box_count_4_A, box_count_4_B, box_count_8_A, box_count_8_B, fractal_1_2_A, fractal_1_2_B, fractal_2_4_A, fractal_2_4_B, fractal_1_4_A, fractal_1_4_B, fractal_4_8_A, fractal_4_8_B, fractal_2_8_A, fractal_2_8_B |
| Camera-independent features (max and min) converted from camera-specific features (A and B) | height[mm]_max, height[mm]_min, number_of_lines_max, number_of_lines_min, pixelwidth[mm]_max, pixelwidth[mm]_min, width[pixel]_max, width[pixel]_min, height[pixel]_max, height[pixel]_min, total_pixels_max, total_pixels_min, area[mm ²]_max, area[mm ²]_min, perimeter[mm]_max, perimeter[mm]_min, box_count_1_max, box_count_1_min, box_count_2_max, box_count_2_min, box_count_4_max, box_count_4_min, box_count_8_max, box_count_8_min, fractal_1_2_max, fractal_1_2_min, fractal_2_4_max, fractal_2_4_min, fractal_1_4_max, fractal_1_4_min, fractal_4_8_max, fractal_4_8_min, fractal_2_8_max, fractal_2_8_min |
| Other features (not used in analysis and classification) | time |

The 2DVD software computes the volume and equivolumetric diameter based on three-dimensional shape reconstructed from two orthogonal projections. The particle shadows in the upper light sheet are matched with particle shadows in the lower sheet, and the software obtains the vertical fall velocity and height quantization (`height_of_one_line`) from the falling time through the planes separated 6.2 mm vertically at the line-scan rate of 34.1 kHz. The number of lines scanned by each camera is the height of the particle. The light sheet of 10 cm is mapped onto 512 pixels in the line-scan camera, and the horizontal resolution of pixel (`pixelwidth`) is about 0.2 mm. The longest scan line is the particle width. The area of each particle was computed by multiplying total number of pixels (`total_pixels`), `height_of_one_line` and `pixelwidth`. We got the boundary of particle shape and computed the particle perimeter.

Camera-specific features are important since they contain various information obtained by 2DVD. However, it is not sufficient to use them directly in the analysis and classification. When we use machine learning algorithms listed in section 3.3, the same type of features obtained by cameras A and B (e.g. `perimeter[mm]_A` and `perimeter[mm]_B`) are also treated as simply different and independent ones. To overcome this problem, we added extra features that are the result of integrating camera-specific features by calculating maximum and minimum values (Figure 3.4). For example, if `perimeter[mm]_A > perimeter[mm]_B`, then `perimeter[mm]_max = perimeter[mm]_A` and `perimeter[mm]_min = perimeter[mm]_B`. In a sense, it is a sorting operation of values from two cameras and if a feature is mainly characterized by large (small) values of it, the integrated feature of its maximum (minimum) will have strong power in the analysis and classification of particles.

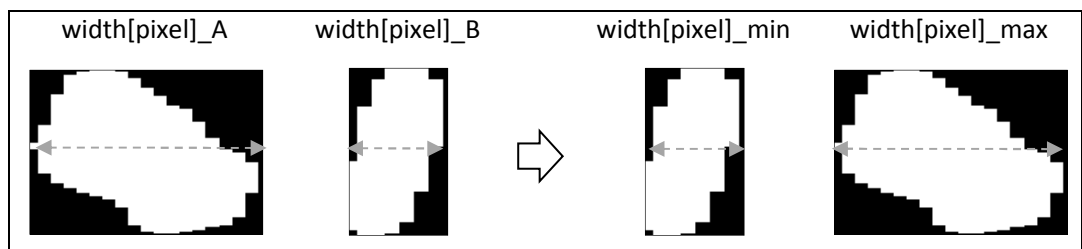
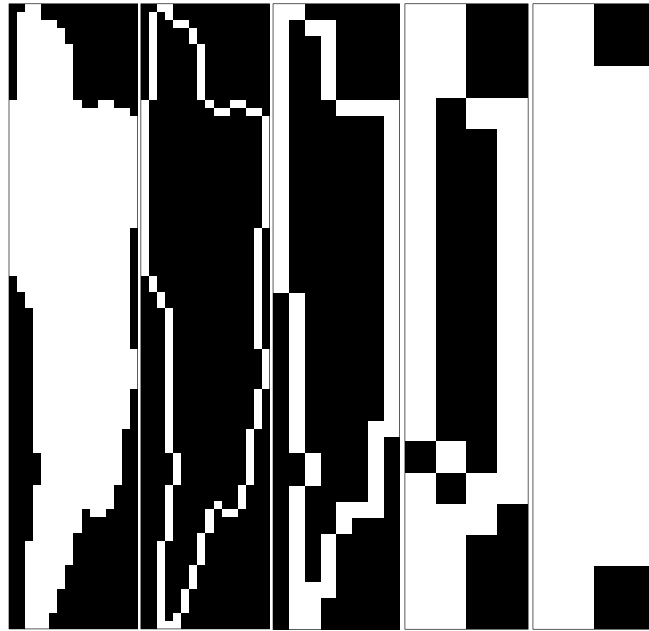


Figure 3.4 Integration of camera-specific features into max and min values.

3.2.2 *Fractal-related features*

Perimeter is a feature that reflects two different characteristics of particle, that is, size and complexity of shape. In this study, we introduced fractal-related features also related to complexity of shape.

Fractal geometry provides a mathematical model for many complex objects with property of self-similarity found in nature[23][24][25]. Fractal dimension is a useful feature for shape classification. The snowflake formation modeled by fractal dimension, was proposed for improvement estimates of snowfall retrieval by radar remote sensing [26][27]. This study uses the box-counting method, which is one of the frequently used techniques to estimate the fractal dimension also known as Minkowski dimension [28][29][30]. First, the smallest number of box shaped elements covering the particle boundary is counted (Figure 3.5). Next, the obtained amount of covering elements is log-log plotted versus the reciprocal of the element size (Figure 3.6). Finally, the box dimension estimate is taken from the monotonically rising linear slope.



(a)



(b)

Figure 3.5 Example of covering results from the box-counting method. (a) Snowflake by camera A; raw image by 2DVD (leftmost), boundary covered by boxes of size 1, 2, 4, and 8. (b) Snowflake by camera B.

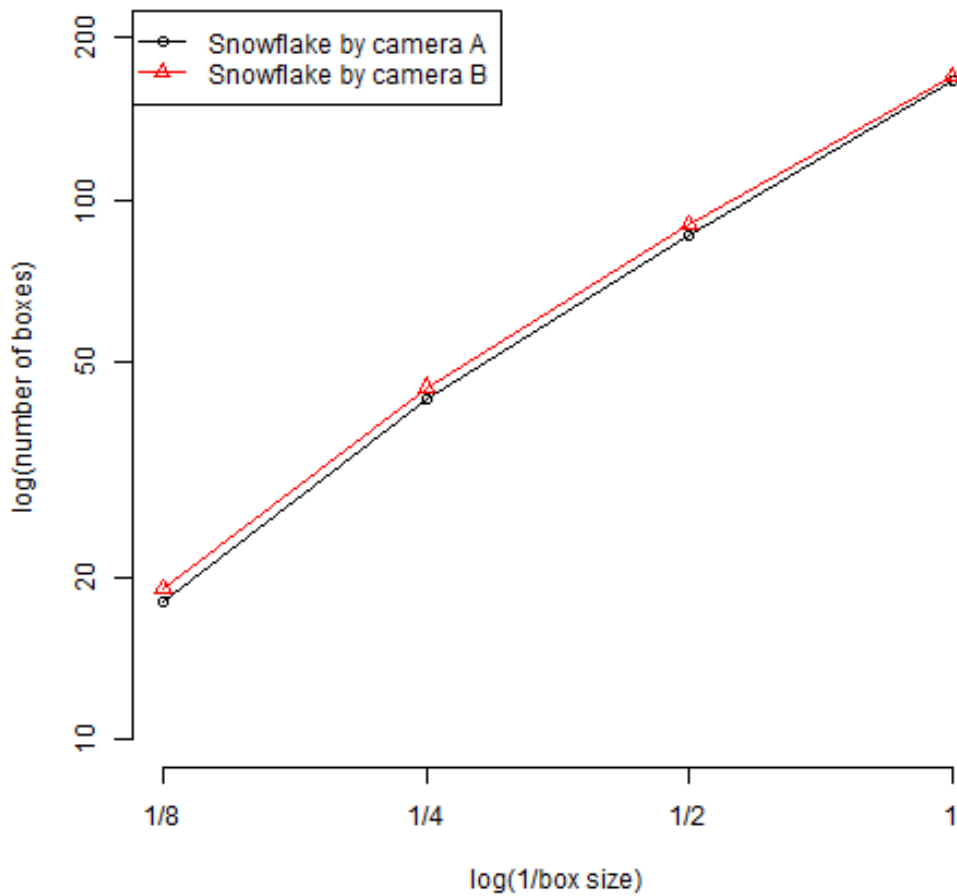


Figure 3.6 The log-log plot of the box-counting method.

3.2.3 Human annotation

Total number of particles in our dataset is 16,010, that is, it consists of 16,010 feature vectors with the features listed in Table 3.1. To conduct meaningful analysis and evaluation of classification performance, we randomly sampled 1,600 feature vectors and annotated them manually. Before annotation, five categories were prepared: *snowflake*, *snowflake-like*, *intermediate*, *graupel-like*, and *graupel*. Additionally, if one of two images for a particle matched one of the following rules, it was automatically annotated as *warning* and filtered out before random sampling since it can be regarded as outlier or erroneous data.

- *equivolumetric_diameter*[mm] is less than 0.2.
- *vertical_fall_velocity*[m/s] is greater than 4.

- width[pixel] / height[pixel] is less than 1/3 or greater than 3.
- The horizontal position of the particle in the raw image is left-end and over 50% of left edge of the particle image is occupied by black pixel (i.e. it is strongly suspected that the particle passed by the left end of a camera and whole image of it was not taken by 2DVD).

The numbers of annotated samples are shown in Table 3.2. According to these annotations, the datasets shown in Table 3.3 are used for analysis and classification in chapter 4.

Table 3.2 The number of samples after annotation.

| Annotation | The number of particles |
|----------------|-------------------------|
| snowflake | 559 |
| snowflake-like | 111 |
| intermediate | 39 |
| graupel-like | 144 |
| graupel | 747 |
| warning | 2,118 |
| not annotated | 12,292 |

Table 3.3 Datasets according to annotation.

| Dataset | Annotation | The number of particles |
|--------------|--|-------------------------|
| whole | snowflake, snowflake-like, intermediate, graupel-like, graupel, warning, not annotated | 16,010 |
| no-warning | snowflake, snowflake-like, intermediate, graupel-like, graupel, not annotated | 13,892 |
| warning-only | warning | 2,118 |
| 5-classes | snowflake, snowflake-like, intermediate, graupel-like, graupel | 1,600 |
| 2-classes | snowflake, graupel | 1,306 |

3.3 Algorithms

In this section, the algorithms we used for analysis and classification are being described.

3.3.1 Normalization

A feature vector consists of two or more feature values for features. However, it is problematic to use the original values for machine learning because in general, value distribution can differ from feature to feature. Therefore, it is popular to normalize the original values of feature vectors so that all the features have the same average and variance. In this study, we normalized our dataset with average = 0 and variance = 1 for each feature before the analysis and classification.

3.3.2 Pearson's correlation coefficient

To see the direct and pairwise relationship between every pair of features, we calculated Pearson's correlation coefficient. If its value is near to 1, two features are quite similar. It is one of the most basic feature analysis methods. In addition, it is known that, removing one of two similar and redundant features may lead to better performance of classification, regression, clustering, etc.

3.3.3 Principal component analysis (PCA)

Among various unsupervised learning algorithms, PCA might be the most popular one[31][32]. Based on the calculation of features' linear combination that maximizes the variance, PCA converts the original feature space into the space of principal components (PCs). After PCA, all the PCs are ordered as PC1, PC2, ... and it is believed that PC1 is the strongest feature for characterizing the feature vectors, PC2 is secondly strong, and so on. Due to this effect of PCA, it is broadly used for different purposes. As the basic analysis of original features, coefficient of each feature in the linear combination formula for some important PCs like PC1 is evaluated. In this study, it may reflect the importance of the feature to characterize and classify snowflakes and graupels.

3.3.4 Support vector machine (SVM)

SVM was first developed by Vladimir Vapnik [33]. Due to its applicability and high-performance, it is one of the most popular machine learning algorithms today.[34][35] Among various variants and implementations of SVM, we used `ksvm` function implemented in `kernlab` package for R. Regarding the choice of kernel, the default one (Radial Basis Function kernel, also known as Gaussian kernel) was adopted. A hyper-parameter “sigma” for this kernel is being automatically optimized by `ksvm`.

3.3.5 Cross-validation

To evaluate the performance of predicting the class label (i.e. snowflake or graupel) of unseen samples (i.e. unseen particles), it is popular to conduct cross-validation. In this study, we adopted 10-fold cross-validation that randomly divides given dataset into 10 and perform learning and prediction 10 times by changing 10% of dataset for test (rest of 90% is used for training). One problem about this kind of cross-validation is that the evaluated performance is affected by the result of random division and different performances are achieved in every evaluation. To solve this problem, we repeated 10-fold cross-validation 100 times and averaged the accuracy.

Chapter 4

Experimental Results and Discussion

In this chapter, four different types of particle analyses are conducted. Experimental results proved that our system is highly accurate and can be a powerful tool for meteorological analysis.

4.1 Feature analysis by Pearson's correlation coefficient

Figure 4.1 illustrates the result of correlation analysis on all feature pairs. It can be summarized as follows:

- Box-count features (i.e. features about the number of boxes) are highly similar to each other. In contrast, fractal features are dissimilar to each other.
- Some of other features are similar to each other (i.e. height and perimeter features). It indicates that redundant features like box-count may exist also in these other features.
- About the difference between camera-specific features ((b), (c), and (d)) and camera-independent features ((e), (f), and (g)) calculated from them, fractal features (d) and (g) showed clear difference. In other words, calculation of max and min was meaningful at least for fractals.

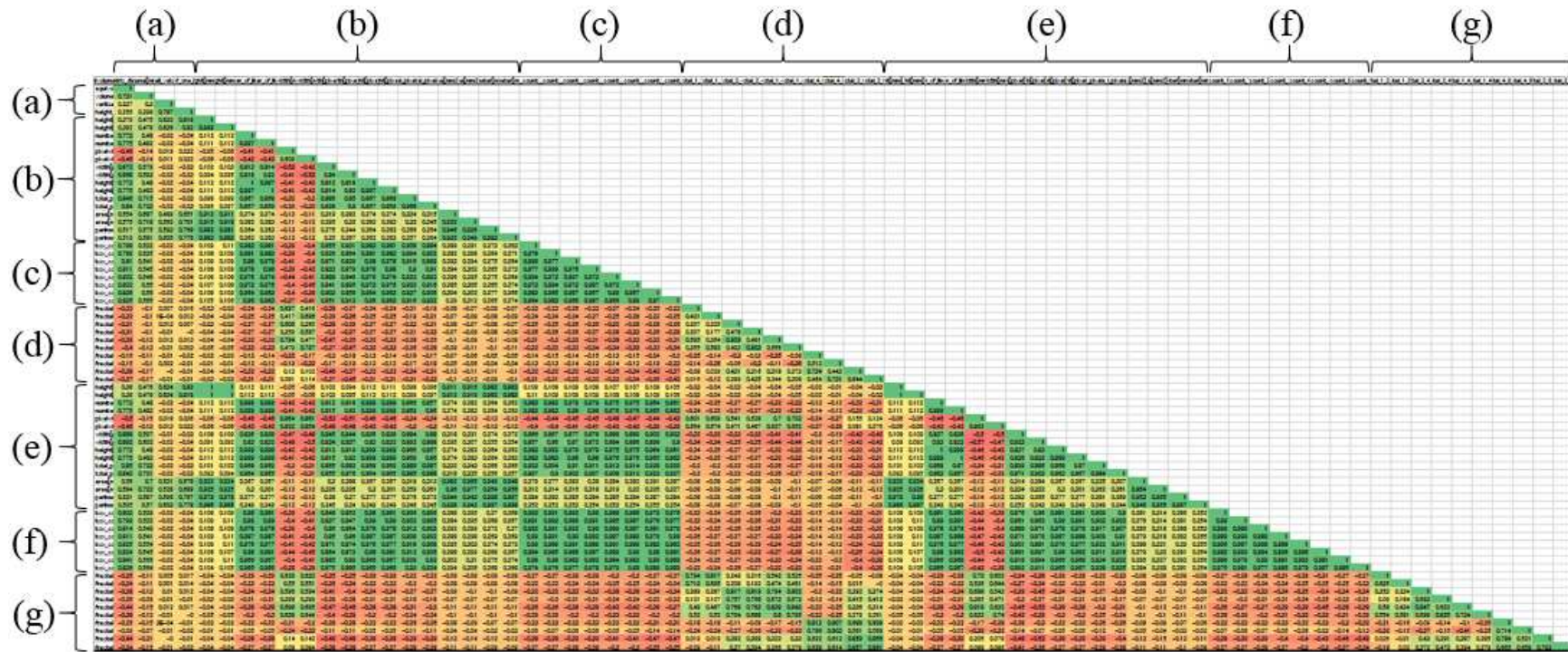


Figure 4.1 Correlation analysis of features. Green (red) color corresponds to high (low) value. (a) Camera-independent features, (b) Camera-specific features, (c) box-count features, (d) fractal features, (e)(f)(g) max and min of (b)(c)(d).

4.2 Feature analysis by PCA

Figures 4.2-4.4 illustrate the similarity among the principal components 1-3 in four datasets (except “warning-only”). In each figure, features are sorted in descending order of principal component of whole dataset. Top 10 important features in each dataset and PC are shown in Tables 4.1-4.3. From these figures and tables, it can be clearly seen that:

- PC1s of these datasets are similar to each other (Figure 4.2). Most of the important features in PC1 are occupied by box-count features (Table 4.1).
- PC2 of the dataset “whole” is quite dissimilar to others (Figure 4.3) and the difference is caused by the inclusion of “warning-only”. In other words, after filtering errors, PC2 is more or less the same in each dataset. About top 10 features of PC1 of “warning-only” (Table 4.1), it is convincing that most of them are occupied by size-related features (height, perimeter, area, etc.) because many of the particles in this dataset were removed from “whole” dataset due to their strange size. About PC2s of the datasets “no-warning”, “5-classes”, and “2-classes”, some of the fractal features occupy top 4 important features (Table 4.2).
- In Figure 4.4, PC3s of the datasets “5-classes” and “2-classes” are quite dissimilar (correlation between them is -0.97). Since in “2-classes”, ambiguous particles annotated as “snowflake-like”, “intermediate”, or “graupel-like” are removed from “5-classes”, it can be interpreted that PC3 of “5-classes” is highly affected by the characteristics of such ambiguous particles.

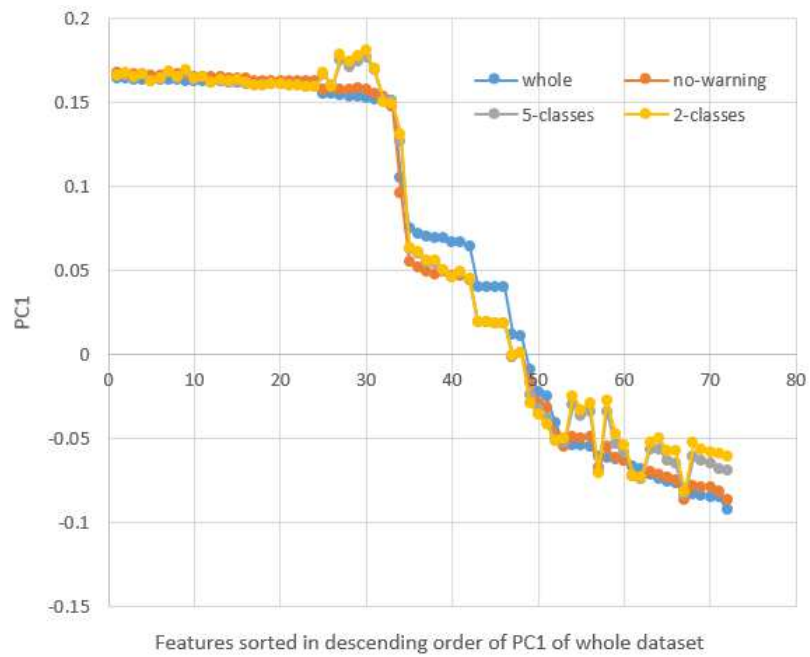


Figure 4.2 PC1 of the datasets except “warning-only”.

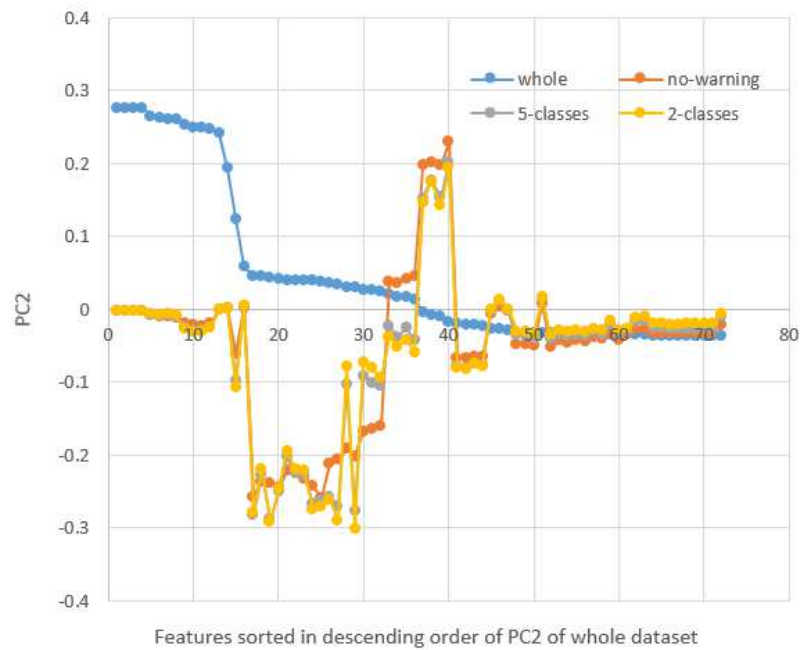


Figure 4.3 PC2 of the datasets except “warning-only”.

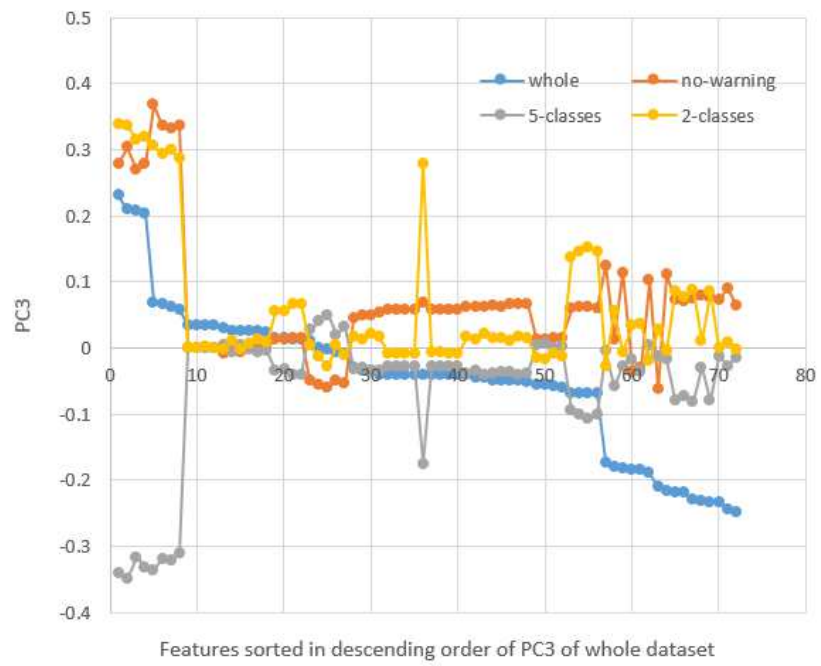


Figure 4.4 PC3 of the datasets except “warning-only”.

Table 4.1 Top 10 features in descending order of PC1 values.

| rank | whole | no-warning | 5-classes | 2-classes | warning-only |
|------|-----------------|-----------------|------------------|------------------|-------------------|
| 1 | box_count_4_min | box_count_4_min | total_pixels_B | total_pixels_B | height[mm]_min |
| 2 | box_count_8_max | box_count_8_min | total_pixels_max | total_pixels_max | height[mm]_B |
| 3 | box_count_4_max | box_count_8_max | total_pixels_min | total_pixels_min | height[mm]_max |
| 4 | box_count_4_B | box_count_8_B | total_pixels_A | total_pixels_A | height[mm]_A |
| 5 | box_count_4_A | box_count_4_B | width[pixel]_B | width[pixel]_B | perimeter[mm]_min |
| 6 | box_count_2_min | box_count_4_max | box_count_8_B | box_count_8_B | perimeter[mm]_B |
| 7 | box_count_8_min | box_count_8_A | box_count_8_min | box_count_8_min | perimeter[mm]_A |
| 8 | box_count_8_A | box_count_2_min | box_count_4_B | width[pixel]_max | perimeter[mm]_max |
| 9 | box_count_8_B | box_count_4_A | width[pixel]_max | box_count_8_max | area[mm2]_max |
| 10 | box_count_2_max | box_count_2_B | box_count_8_max | box_count_4_B | area[mm2]_min |
| 10 | box_count_2_max | box_count_2_B | box_count_8_max | box_count_4_B | area[mm2]_min |

Table 4.2 Top 10 features in descending order of PC2 values.

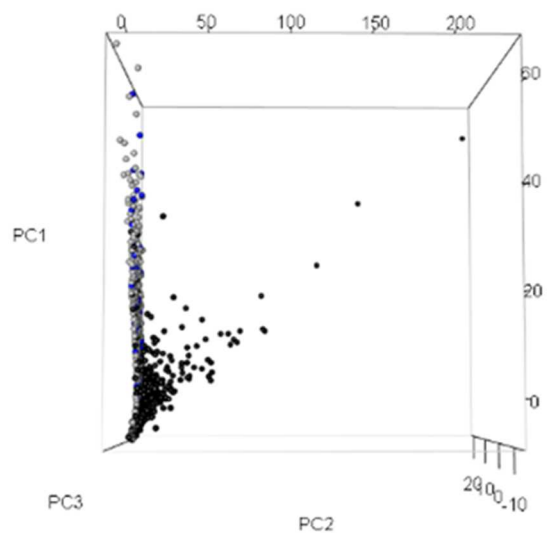
| rank | whole | no-warning | 5-classes | 2-classes | warning-only |
|------|-------------------|------------------|---------------------------------|---------------------------------|------------------------|
| 1 | height[mm]_B | fractal_4_8_min | fractal_4_8_min | fractal_4_8_min | pixelwidth[mm]_min |
| 2 | height[mm]_min | fractal_4_8_B | fractal_4_8_B | fractal_4_8_B | pixelwidth[mm]_B |
| 3 | height[mm]_max | fractal_4_8_A | fractal_4_8_A | fractal_4_8_max | pixelwidth[mm]_max |
| 4 | height[mm]_A | fractal_4_8_max | fractal_4_8_max | fractal_4_8_A | pixelwidth[mm]_A |
| 5 | perimeter[mm]_min | fractal_2_8_min | width[pixel]_min | width[pixel]_min | fractal_1_2_max |
| 6 | perimeter[mm]_B | fractal_2_8_B | width[pixel]_A | width[pixel]_A | fractal_1_2_min |
| 7 | perimeter[mm]_A | fractal_2_8_max | equivolumetric_ diameter[mm] | equivolumetric_ diameter[mm] | fractal_1_2_B |
| 8 | perimeter[mm]_max | fractal_2_8_A | vertical_fall_ velocity[m/s] | vertical_fall_ velocity[m/s] | fractal_1_2_A |
| 9 | area[mm2]_max | width[pixel]_min | height_of_one_ line[mm] | width[pixel]_B | height_of_one_line[mm] |
| 10 | area[mm2]_B | width[pixel]_A | width[pixel]_B | height_of_one_ line[mm] | fractal_1_4_max |

Table 4.3 Top 10 features in descending order of PC3 values.

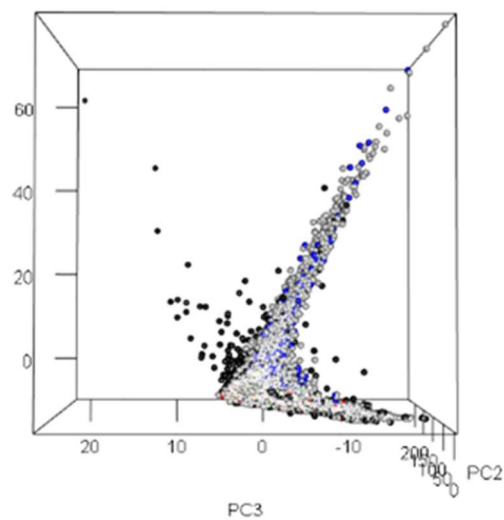
| rank | whole | no-warning | 5-classes | 2-classes | warning-only |
|------|-----------------|-----------------|---------------------------------|-----------------|--------------------|
| 1 | fractal_4_8_min | fractal_2_8_min | width[pixel]_A | fractal_4_8_min | pixelwidth[mm]_min |
| 2 | fractal_4_8_max | fractal_2_8_A | width[pixel]_min | fractal_4_8_max | pixelwidth[mm]_B |
| 3 | fractal_4_8_B | fractal_2_8_B | width[pixel]_max | fractal_4_8_A | pixelwidth[mm]_A |
| 4 | fractal_4_8_A | fractal_2_8_max | equivolumetric_ diameter[mm] | fractal_4_8_B | volume[mm3] |
| 5 | fractal_2_8_min | fractal_4_8_max | width[pixel]_B | fractal_2_8_min | width[pixel]_max |
| 6 | fractal_2_8_B | fractal_4_8_min | box_count_8_A | fractal_2_8_max | width[pixel]_A |
| 7 | fractal_2_8_max | fractal_4_8_A | height_of_one_ line[mm] | fractal_2_8_B | box_count_8_max |
| 8 | fractal_2_8_A | fractal_4_8_B | vertical_fall_ velocity[m/s] | fractal_2_8_A | box_count_8_A |
| 9 | height[mm]_max | fractal_2_4_min | box_count_8_max | volume[mm3] | total_pixels_max |
| 10 | height[mm]_A | fractal_2_4_B | fractal_2_4_A | total_pixels_B | box_count_8_B |

For visually understanding the sample distribution, we show 3D plots of the datasets. In Figures 4.5-4.9, it can be seen that the distributions of samples in three datasets “no-warning”, “5-classes”, and “2-classes” are almost the same. The 3D plots from three angles for “5-classes” show that, snowflake samples have their own distribution distinguishable from others. In contrast, samples of other annotations (snowflake-like, intermediate, graupel-like, and graupel) are distributed in the plane near to the PC2-PC3. About the L-like distribution of these samples, it is caused by the combined use of camera-specific fractal features (fractal_1_2_A, ..., fractal_2_8_B) and camera-independent fractal features (fractal_1_2_max, ..., fractal_2_8_min). For example, removal of box-count features does not affect to the L-like shape of the distribution, however, removal of camera-specific or camera-independent fractal features makes it ambiguous (Figure 4.10 and 4.11). Although the meaning of the distribution is still unclear, this result suggests that the fractal features could provide more detailed classification of non-snowflake particles.

View from PC1-PC2 plane



View from PC1-PC3 plane



View from PC2-PC3 plane

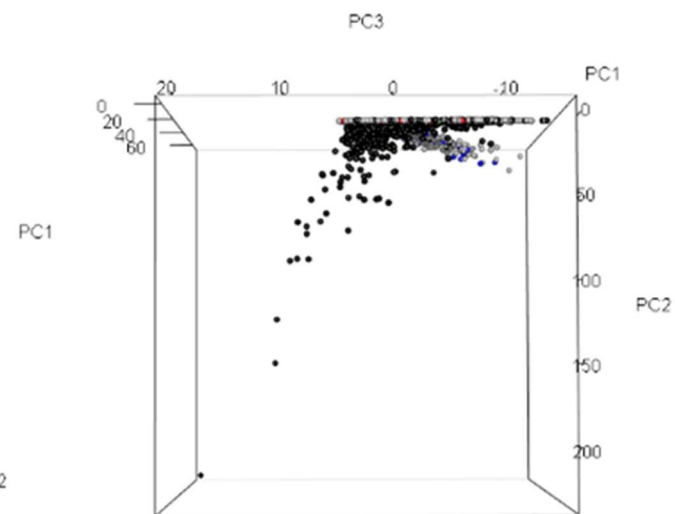
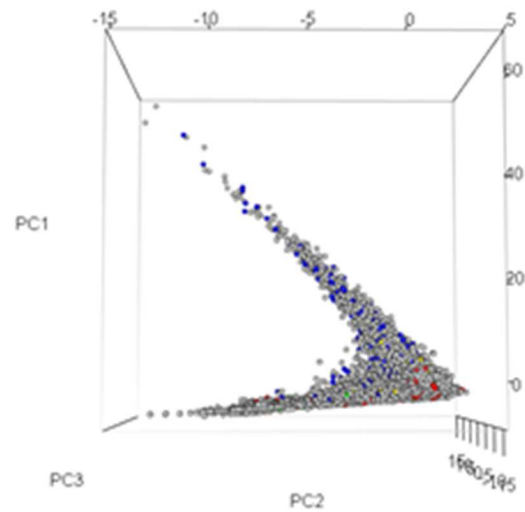
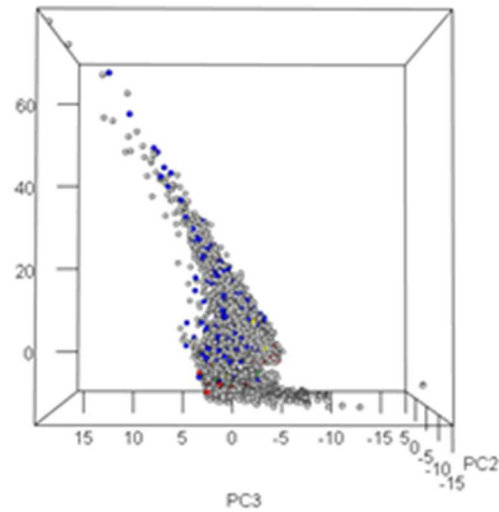


Figure 4.5 3D plots of PC1, PC2, and PC3 in the dataset “all” from three different angles of view. The colors of points (gray, black, blue, green, white, yellow, red) indicate the annotations (not annotated, warning, snowflake, snowflake-like, intermediate, graupel-like, graupel), respectively.

View from PC1-PC2 plane



View from PC1-PC3 plane



View from PC2-PC3 plane

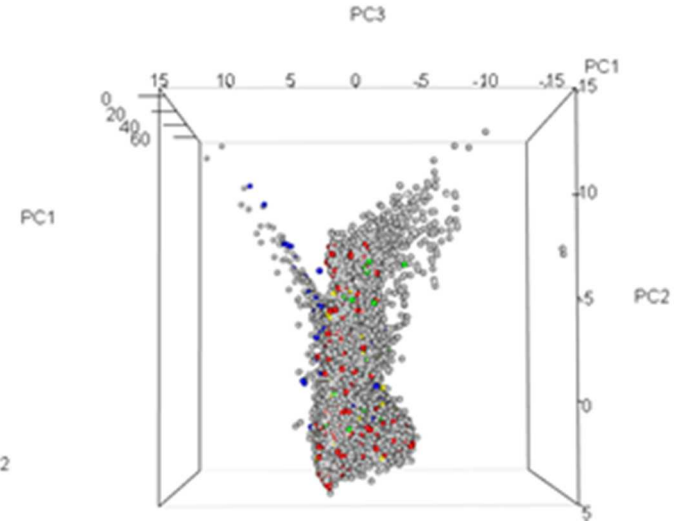
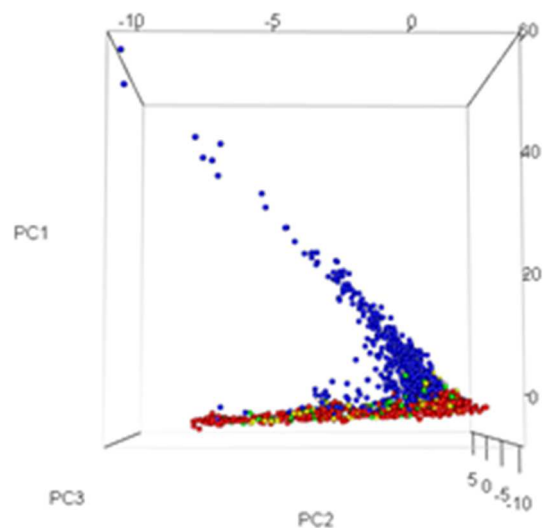
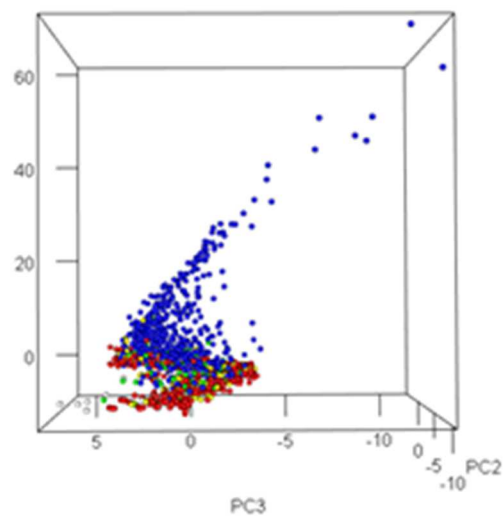


Figure 4.6 3D plots of PC1, PC2, and PC3 in the dataset “no-warning” from three different angles of view. The colors of points (gray, blue, green, white, yellow, red) indicate the annotations (not annotated, snowflake, snowflake-like, intermediate, graupel-like, graupel), respectively.

View from PC1-PC2 plane



View from PC1-PC3 plane



View from PC2-PC3 plane

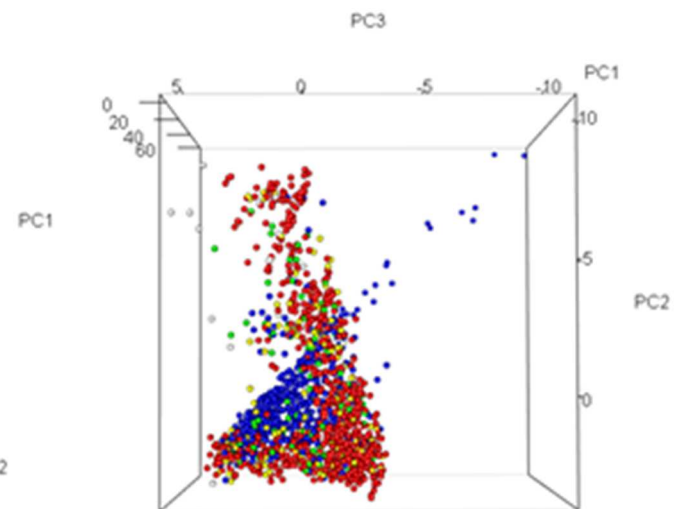
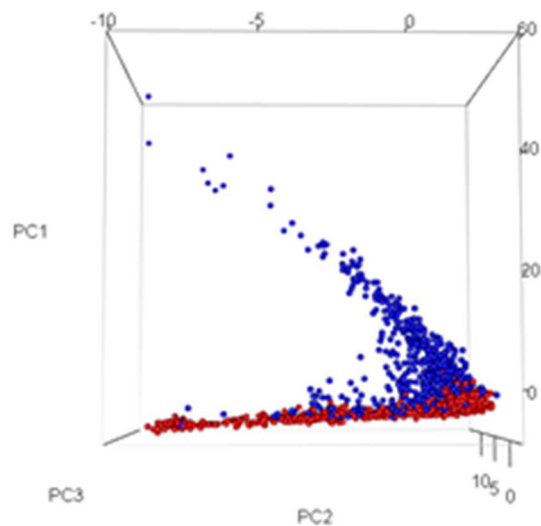
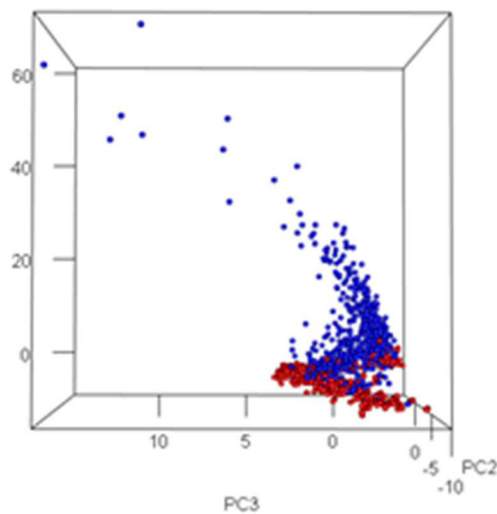


Figure 4.7 3D plots of PC1, PC2, and PC3 in the dataset “5-classes” from three different angles of view. The colors of points (blue, green, white, yellow, red) indicate the annotations (snowflake, snowflake-like, intermediate, graupel-like, graupel), respectively.

View from PC1-PC2 plane



View from PC1-PC3 plane



View from PC2-PC3 plane

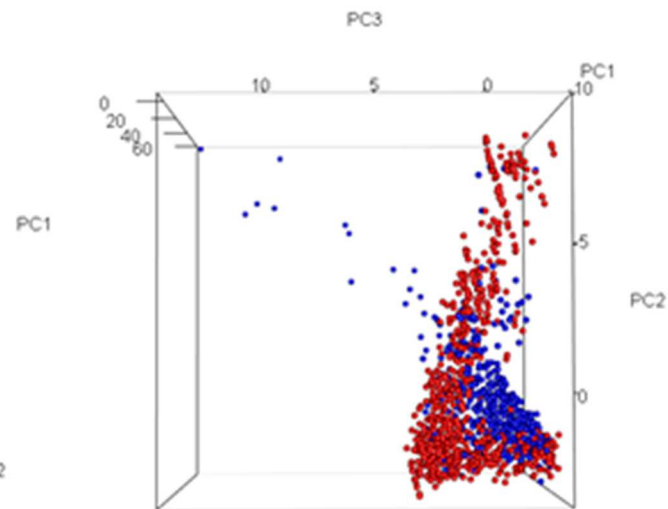
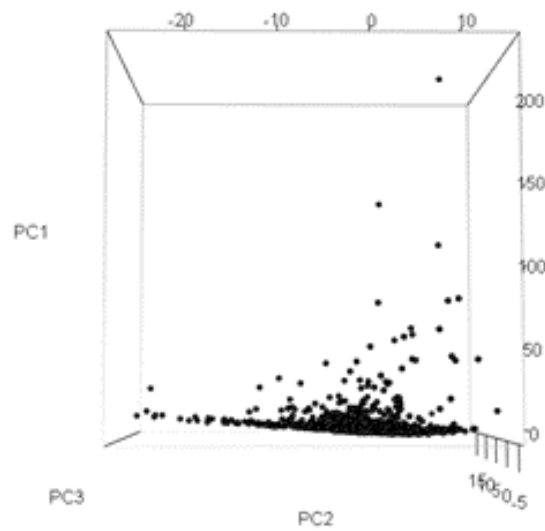
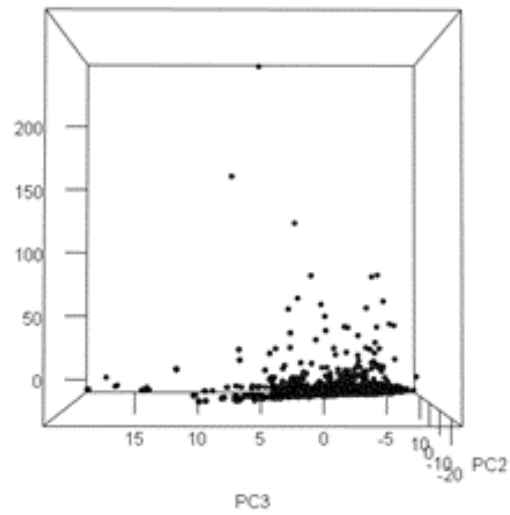


Figure 4.8 3D plots of PC1, PC2, and PC3 in the dataset “2-classes” from three different angles of view. The colors of points (blue, red) indicate the annotations (snowflake, graupel), respectively.

View from PC1-PC2 plane



View from PC1-PC3 plane



View from PC2-PC3 plane

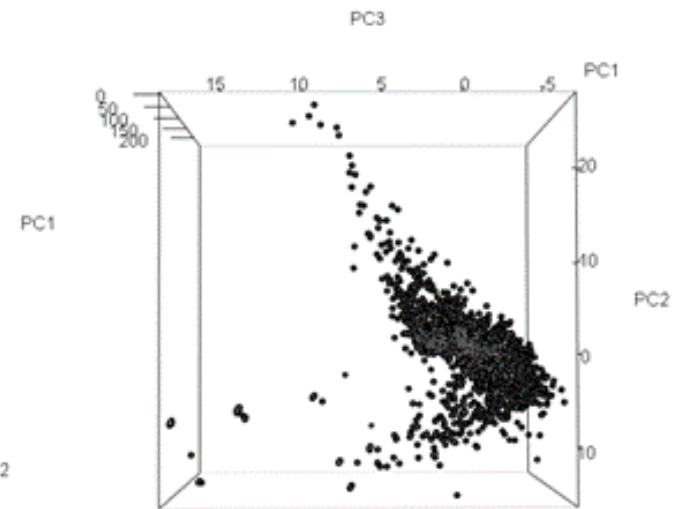


Figure 4.9 3D plots of PC1, PC2, and PC3 in the dataset “warning” from three different angles of view.

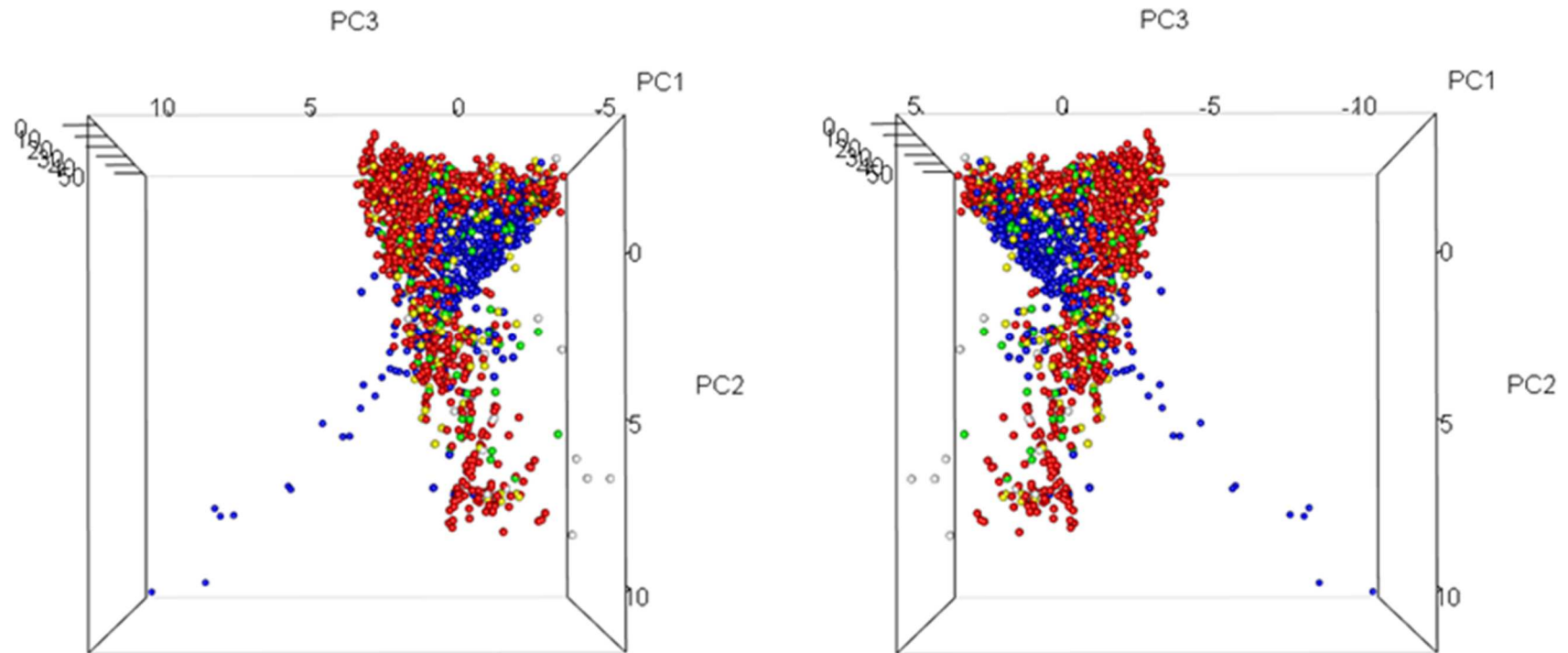


Figure 4.10 3D plots of the dataset “5-classes” without some features. In left and right panels, camera-specific and camera-independent box-count features are removed, respectively.

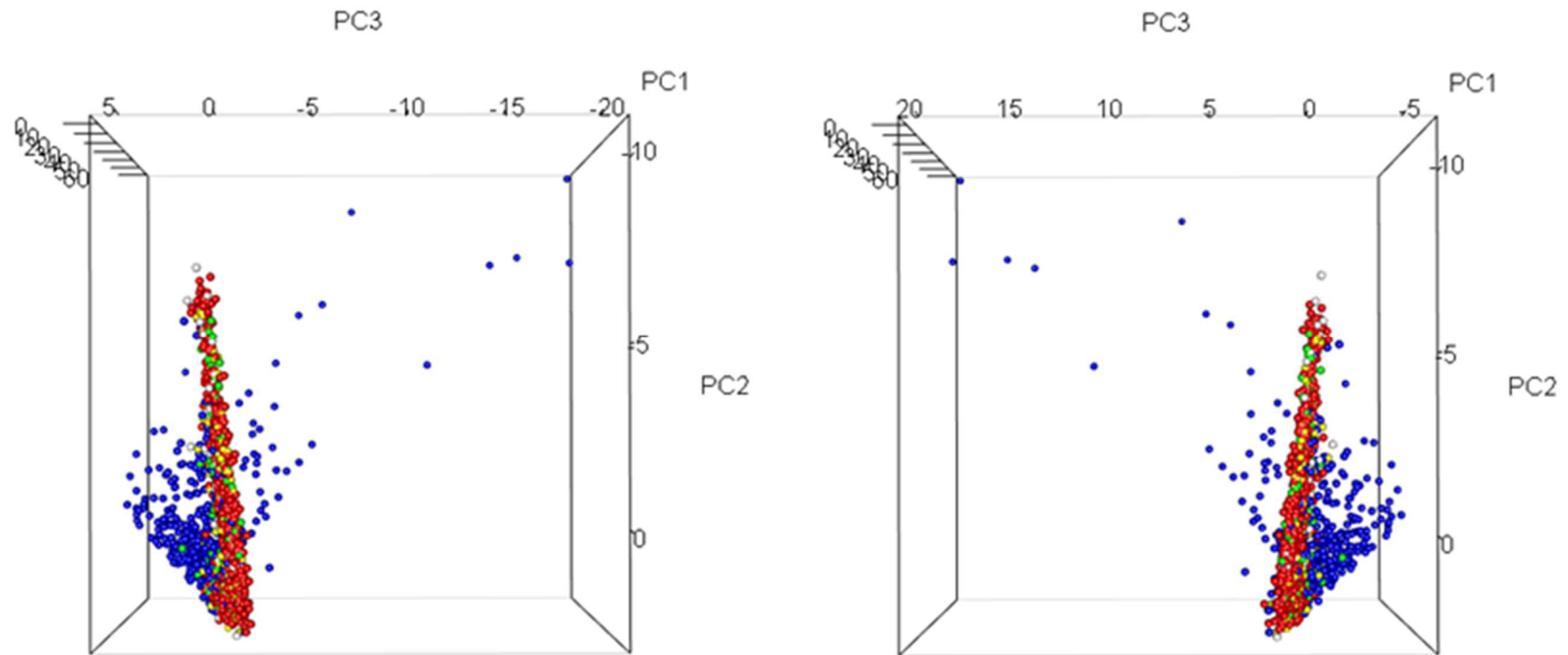


Figure 4.11 3D plots of the dataset “5-classes” without some features. In left and right panels, camera-specific and camera-independent fractal features are removed, respectively.

4.3 Particle classification by SVM

As shown in Table 3.1, 72 features are available for training a statistical model to classify given samples (particles) into snowflakes and graupels. Using the algorithms described in subsections 3.3.4 and 3.3.5, first we evaluated the accuracy of prediction with “2-classes” dataset and all 72 features. The average error of prediction (i.e. 1 - average accuracy) was 0.08263. After converting the 72 features into 72 PCs by PCA, the average error decreased to 0.07191.

Since so many redundant features exist in the 72 features, reduction of feature set by feature selection might decrease the average error of prediction. Although various algorithms have been proposed for fully-automatic feature selection, in this study we initially tried to select a representative feature in each feature group, assuming a feature group consisting of all features with common name prefix. For example, `perimeter[mm]_A`, `perimeter[mm]_B`, `perimeter[mm]_max`, and `perimeter[mm]_min` belong to the same group. In case of box-count and fractal features, numbers in the names were ignored since they are homogenous except the parameters for calculating them. To choose the representative feature in each group, 72 evaluations were performed using only one specific feature in each evaluation. As a result, 14 representative features with the lowest average errors in their groups were selected (Table 4.4). Among them, `box_count_2_max` achieved the best performance (0.1055) as a single feature. It is also notable that the suffixes “_max” and “_min” frequently appear instead of “_A” and “_B”. It indicates that the conversion of camera specific features to camera-independent ones contributed to achieve better classification performance.

Starting from the feature set with all of these 14 features, feature selection by backward elimination was performed. It is an iterative feature selection method which removes a feature in an iteration. If the size of feature set in the iteration i is n_i , all subsets with size $n_i - 1$ are evaluated, and if the elimination of a feature achieved the best improvement of average error, it is removed in the next iteration. As a baseline performance before the 1st iteration, the average error 0.0543 achieved by the feature set with all of these 14 features was used.

In this study, four features were removed through 1st to 4th iterations, and the process of backward elimination stopped since 5th iteration could not achieve any improvement. Using the remaining 10 features, the average error 0.0461 was achieved and it was the best performance of classification in this study¹. Unlike the analysis in section 4.2, this result revealed that fractal features could not contribute to the best performance. In other words, they might be useful for more detailed characterization of various particles, not for just classifying snowflakes and graupels. In contrast, a box-count feature (box_count_2_max) was so important as to the classification by only one feature achieved average error 0.1055 that is nearly 90% accuracy. It is an interesting finding that, although a box-count feature is a by-product of fractal calculation, it is significantly important in the classification of snowflakes and graupels.

¹We conducted t-test on two groups of errors before calculating 0.0465 and 0.0461 in Table 7, but it did not show statistically significant difference (p-value = 0.05153). However, at least it was confirmed that 0.0484 and 0.0465 were significantly different (p-value = 3.815e-14).

Table 4.4 Average errors (i.e. 1 – average accuracy) in the predictions by single feature and multiple features with backward elimination. Before backward elimination, average error of prediction by using all 14 features listed in the first column was 0.0543. In each iteration of backward elimination, if the elimination of a feature decreased (increased) the average error of prediction, it is shown in red (blue) color. The least average error in each column is shown in bold face and the corresponding feature is being not used in the succeeding iterations of backward elimination.

| feature | prediction by single feature | 1 st iteration | 2 nd iteration | 3 rd iteration | 4 th iteration | 5 th iteration |
|-----------------------------|------------------------------|---------------------------|---------------------------|---------------------------|---------------------------|---------------------------|
| box_count_2_max | 0.1055 | 0.0599 | 0.0543 | 0.0481 | 0.0493 | 0.0463 |
| total_pixels_max | 0.1198 | 0.0577 | 0.0538 | 0.0485 | 0.0461 | removed |
| number_of_lines_min | 0.1222 | 0.0549 | 0.0511 | 0.0485 | 0.0480 | 0.0466 |
| height[pixel]_min | 0.1224 | 0.0548 | 0.0513 | 0.0481 | 0.0480 | 0.0467 |
| perimeter[mm]_max | 0.1274 | 0.0683 | 0.0665 | 0.0626 | 0.0654 | 0.0653 |
| width[pixel]_max | 0.1405 | 0.0564 | 0.0509 | 0.0471 | 0.0479 | 0.0476 |
| area[mm2]_max | 0.1886 | 0.0602 | 0.0574 | 0.0495 | 0.0526 | 0.0522 |
| height[mm]_min | 0.1913 | 0.0546 | 0.0531 | 0.0465 | removed | removed |
| equivolumetric_diameter[mm] | 0.2026 | 0.0652 | 0.0622 | 0.0556 | 0.0561 | 0.0573 |
| volume[mm3] | 0.2045 | 0.0567 | 0.0506 | 0.0481 | 0.0486 | 0.0469 |
| fractal_2_8_min | 0.2069 | 0.0520 | 0.0484 | removed | removed | removed |
| pixelwidth[mm]_max | 0.2434 | 0.0517 | removed | removed | removed | removed |
| height_of_one_line [mm] | 0.3449 | 0.0557 | 0.0529 | 0.0509 | 0.0504 | 0.0513 |
| vertical_fall_velocity[m/s] | 0.4261 | 0.0556 | 0.0522 | 0.0503 | 0.0499 | 0.0503 |

4.4 Classification of unlabeled data

In the previous section, we optimized the set of features and achieved low average error (i.e. 0.0461). It means that now we can accurately predict the type of particles without annotation. Hereinafter, we define the dataset “unlabeled” by subtracting “5-classes” from “no-warning”. The size of this dataset is 12,292.

Before the prediction, we trained a model for classifying snowflake and graupel using all data in “2-classes” dataset. The set of features is the same as the best one optimized in the previous section. Using this trained model, we predicted the type of particles in “unlabeled” dataset. As might be expected, the result was similar to the one for “2-classes” (Figure 4.12).

A key question here is whether our system is useful for analyzing time-varying behavior of precipitation. In Figures 4.13-4.21, the result of prediction is shown with a time axis which indicates 600 seconds of observation. Among them, Figures 4.13-4.15 with PC1, PC2, and time clearly illustrated the dynamic change of type and amount of particles. For example, we can see that after 320 seconds, the variety of snowflakes suddenly increased. In contrast, the variety of graupels decreased around 100 seconds and 200 seconds.

Finally, Figure 4.22 and 4.23 show the histograms of snowflakes and graupels. We can easily see that both of snowflakes and graupels are suddenly increased around 260 seconds. After that, graupels rapidly decreased whereas snowflakes kept a level of intensity. Although these figures are based on the result of prediction, they demonstrated that such a fine-grained (i.e. particle-by-particle) prediction of particles can be a powerful tool for understanding behavior of snowfall and meteorological theories behind it.

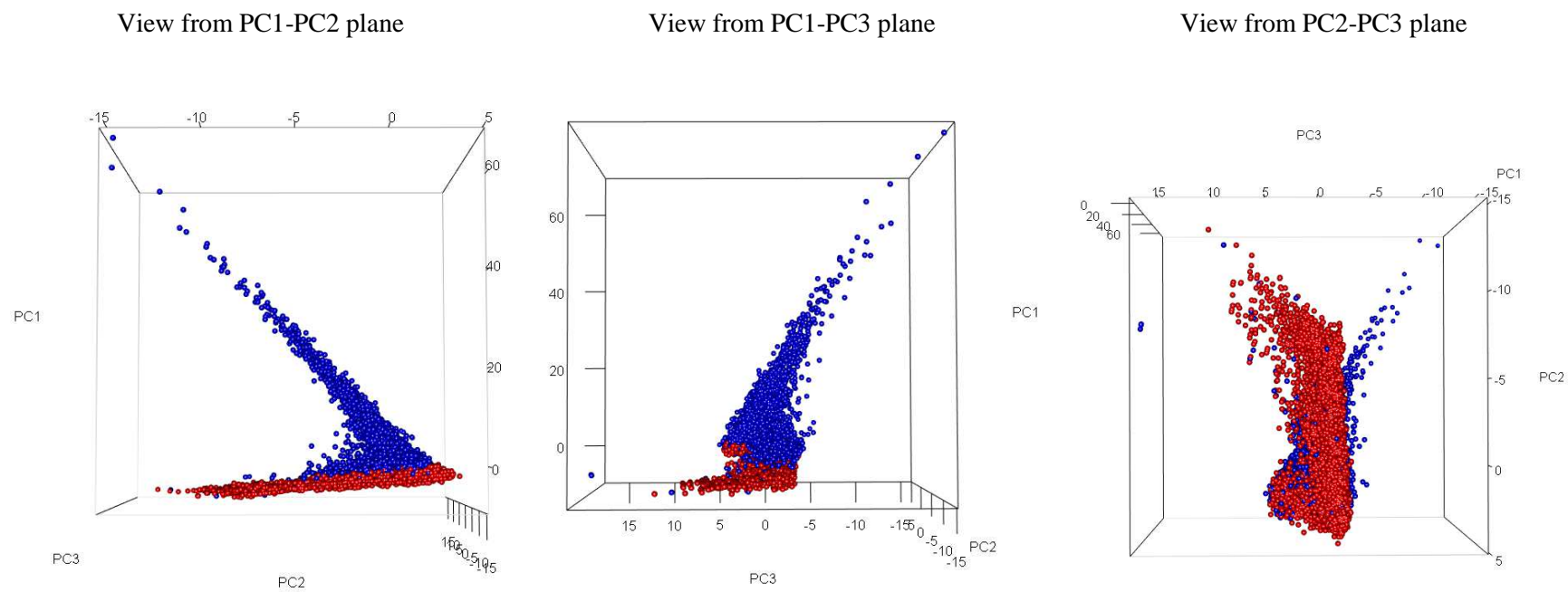


Figure 4.12 3D plots of PC1, PC2, and PC3 in the dataset “unlabeled” from three different angles of view. The colors of points (blue and red) indicate the predicted class labels (snowflake and graupel), respectively.

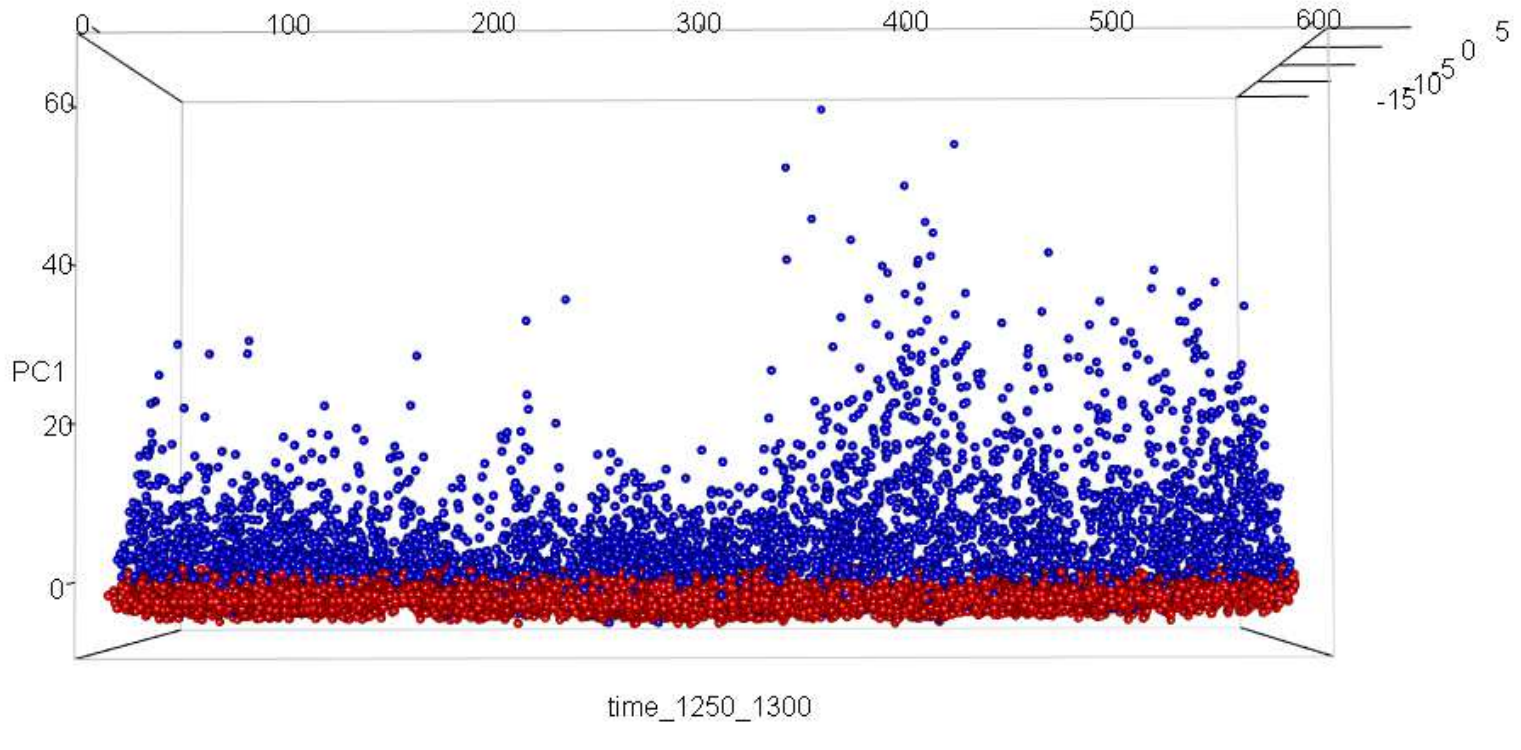


Figure 4.13 3D plots of PC1, PC2, and time in the dataset “unlabeled” with the view from PC1-time plane. The colors of points (blue and red) indicate the predicted class labels (snowflake and graupel), respectively.

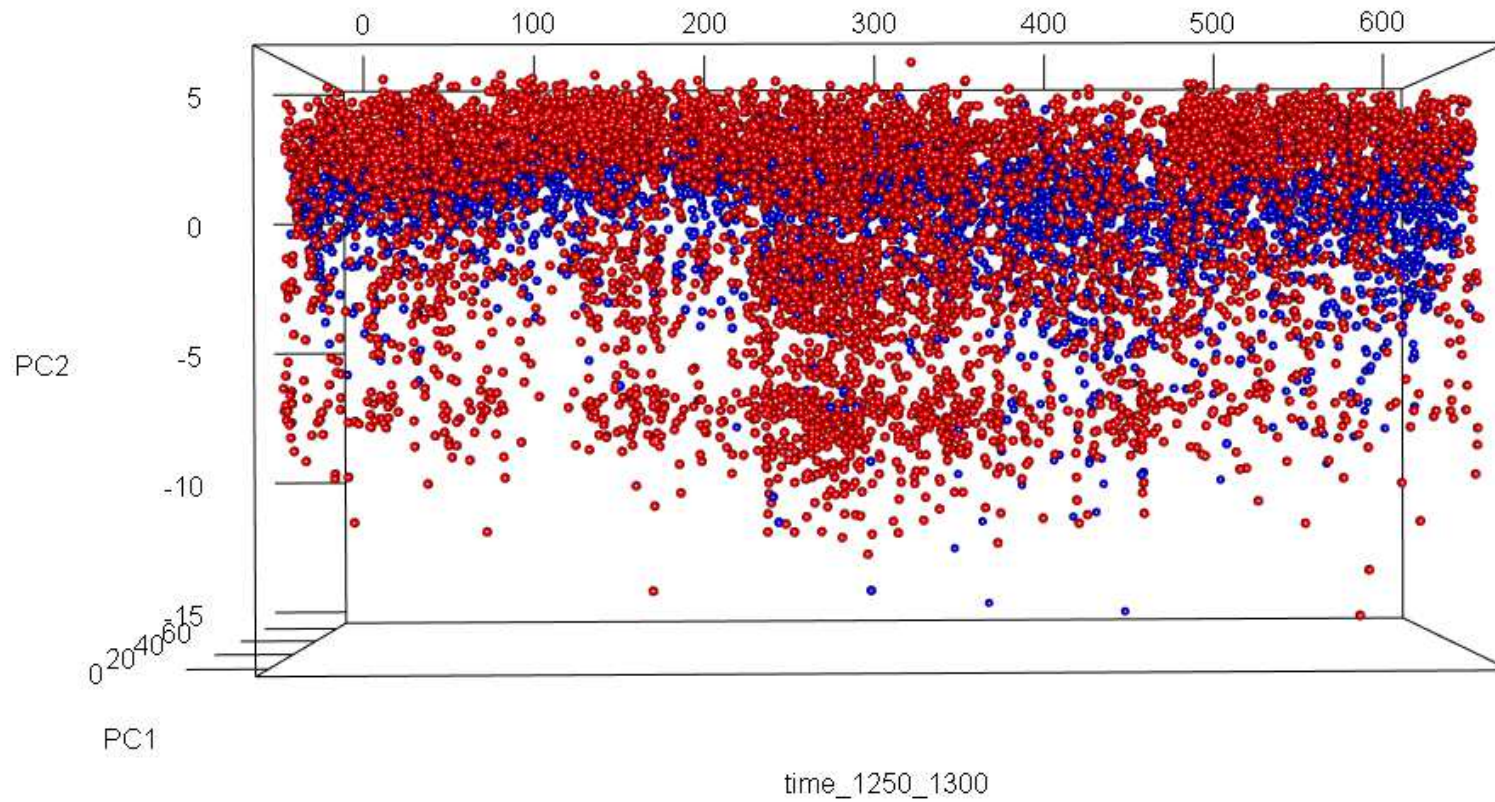


Figure 4.14 3D plots of PC1, PC2, and time in the dataset “unlabeled” with the view from PC2-time plane. The colors of points (blue and red) indicate the predicted class labels (snowflake and graupel), respectively.

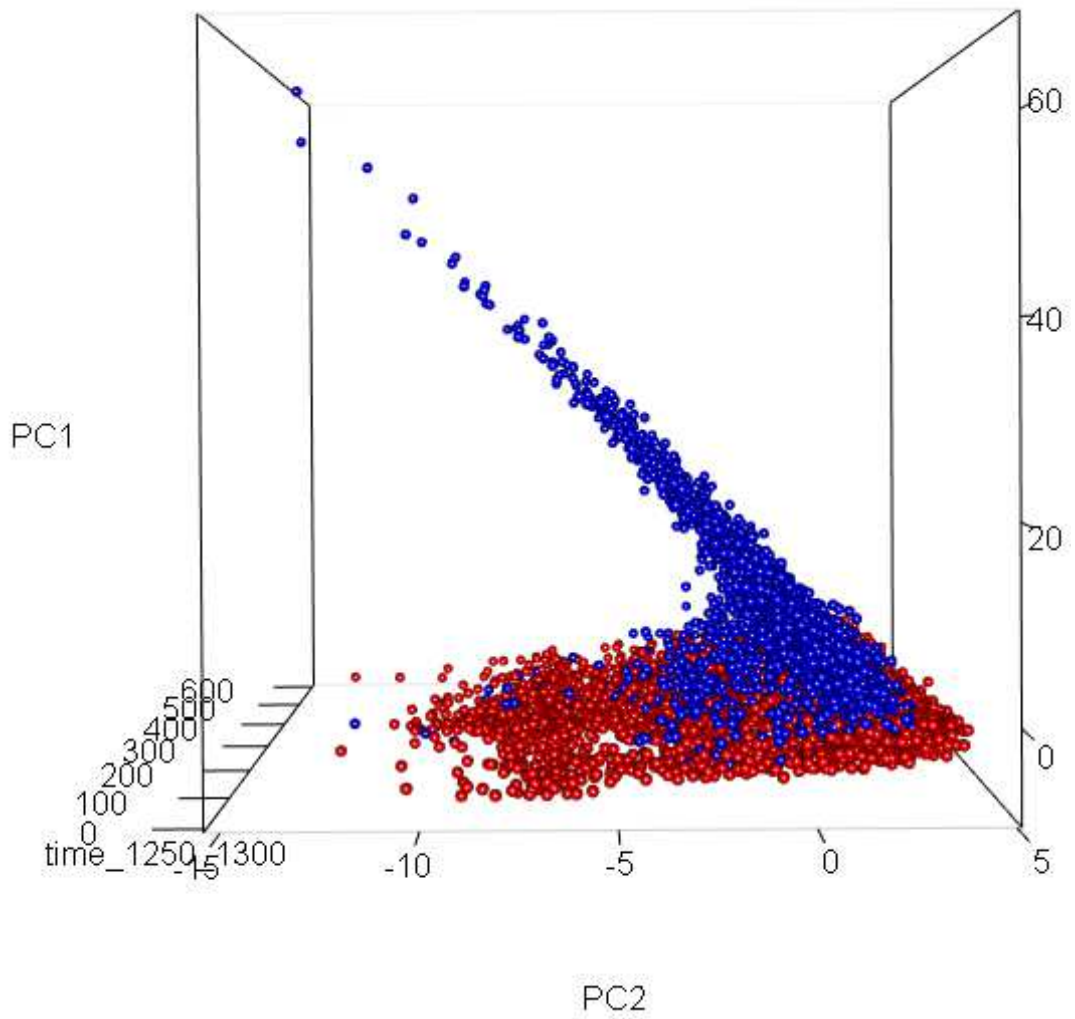


Figure 4.15 3D plots of PC1, PC2, and time in the dataset “unlabeled” with the view from PC1-time plane. The colors of points (blue and red) indicate the predicted class labels (snowflake and graupel), respectively.

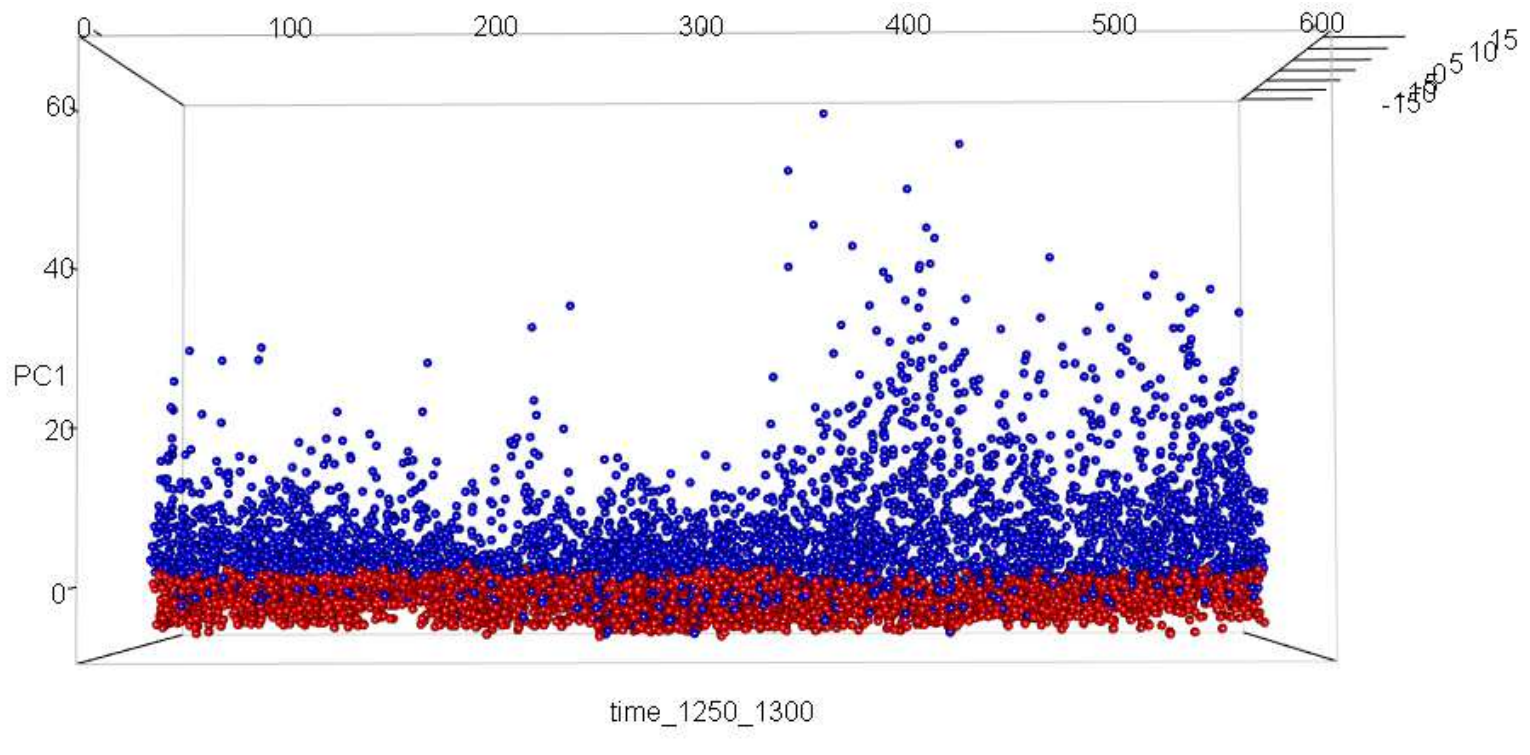


Figure 4.16 3D plots of PC1, PC3, and time in the dataset “unlabeled” with the view from PC1-time plane. The colors of points (blue and red) indicate the predicted class labels (snowflake and graupel), respectively.

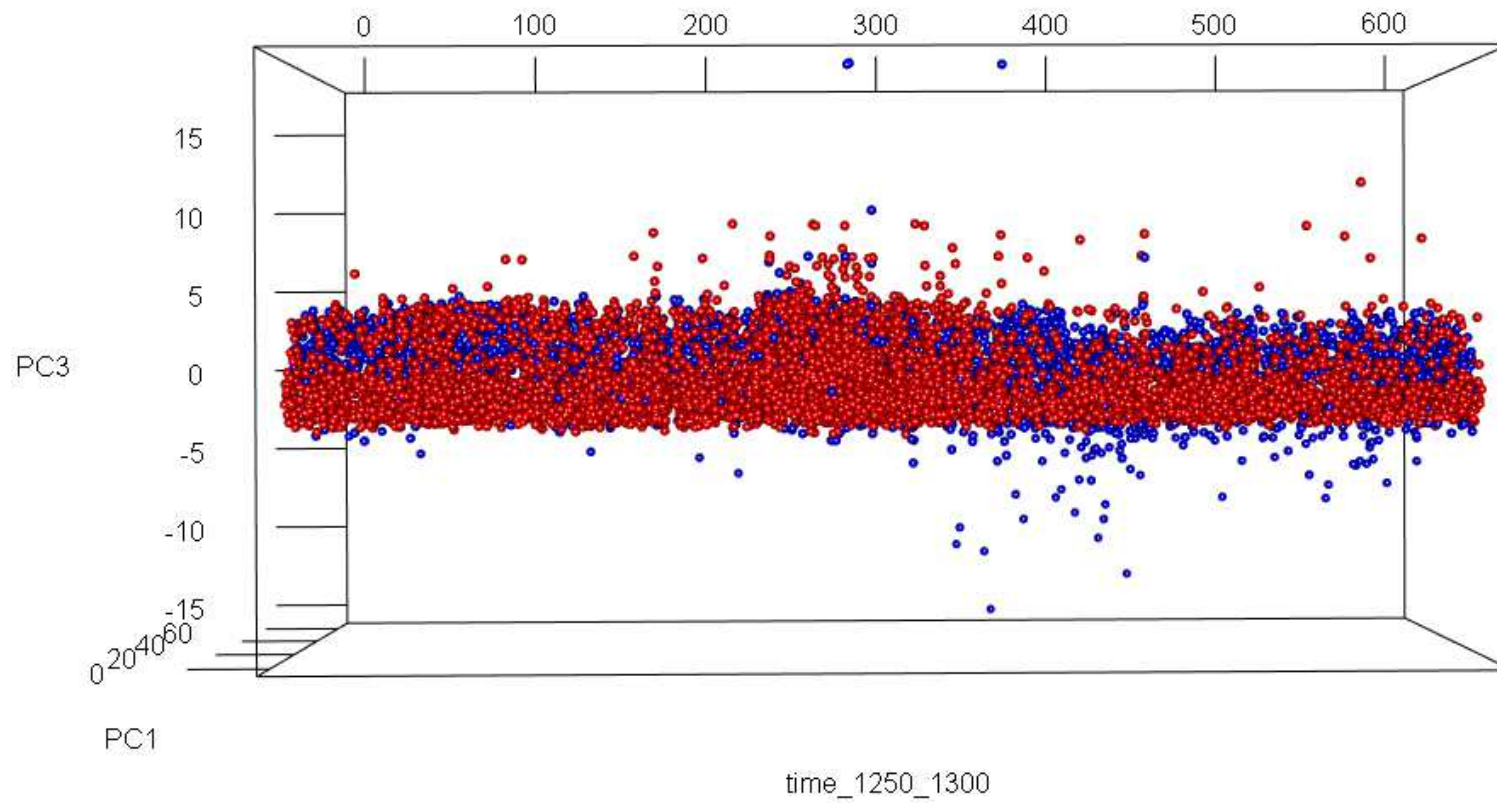


Figure 4.17 3D plots of PC1, PC3, and time in the dataset “unlabeled” with the view from PC3-time plane. The colors of points (blue and red) indicate the predicted class labels (snowflake and graupel), respectively.

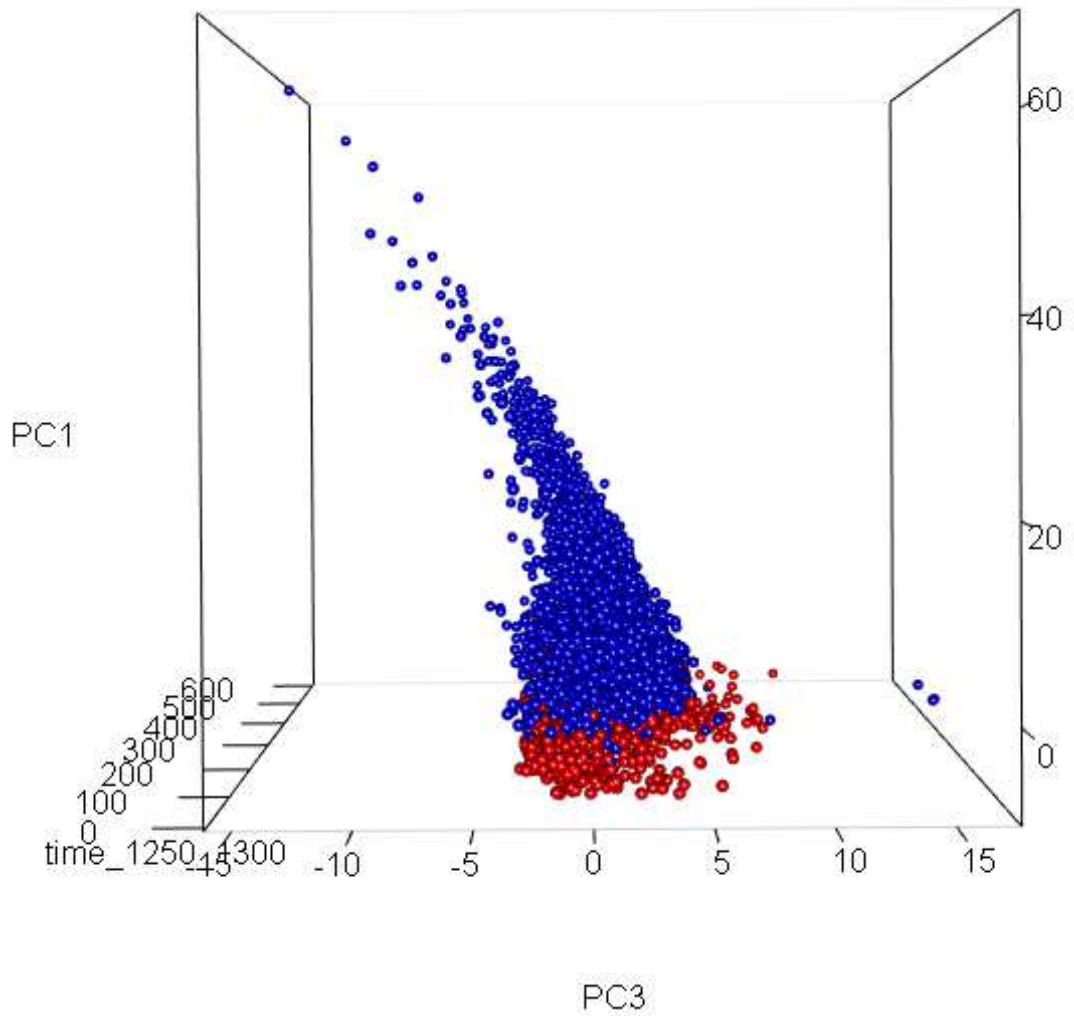


Figure 4.18 3D plots of PC1, PC2, and time in the dataset “unlabeled” with the view from PC1-time plane. The colors of points (blue and red) indicate the predicted class labels (snowflake and graupel), respectively.

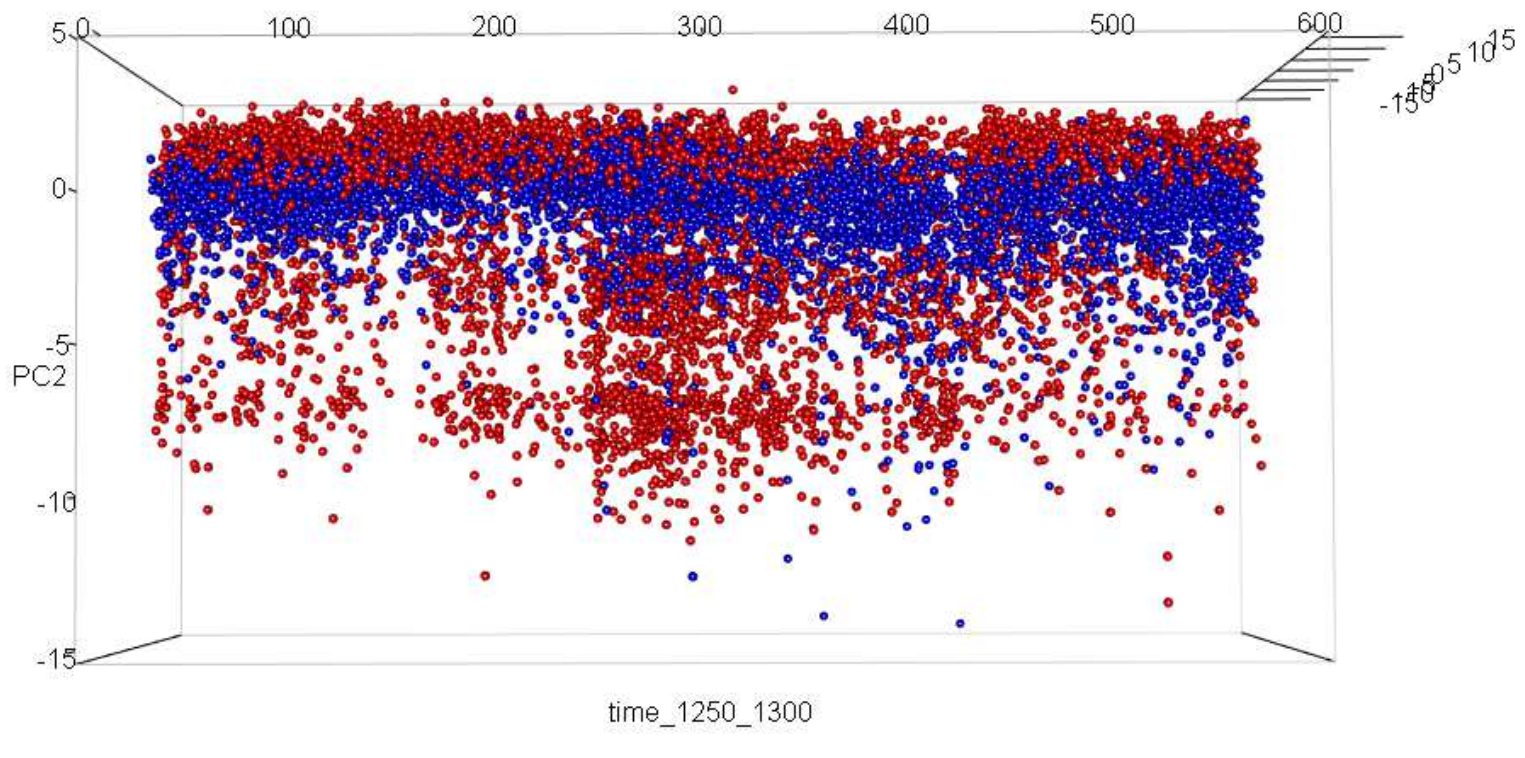


Figure 4.19 3D plots of PC2, PC3, and time in the dataset “unlabeled” with the view from PC2-time plane. The colors of points (blue and red) indicate the predicted class labels (snowflake and graupel), respectively.

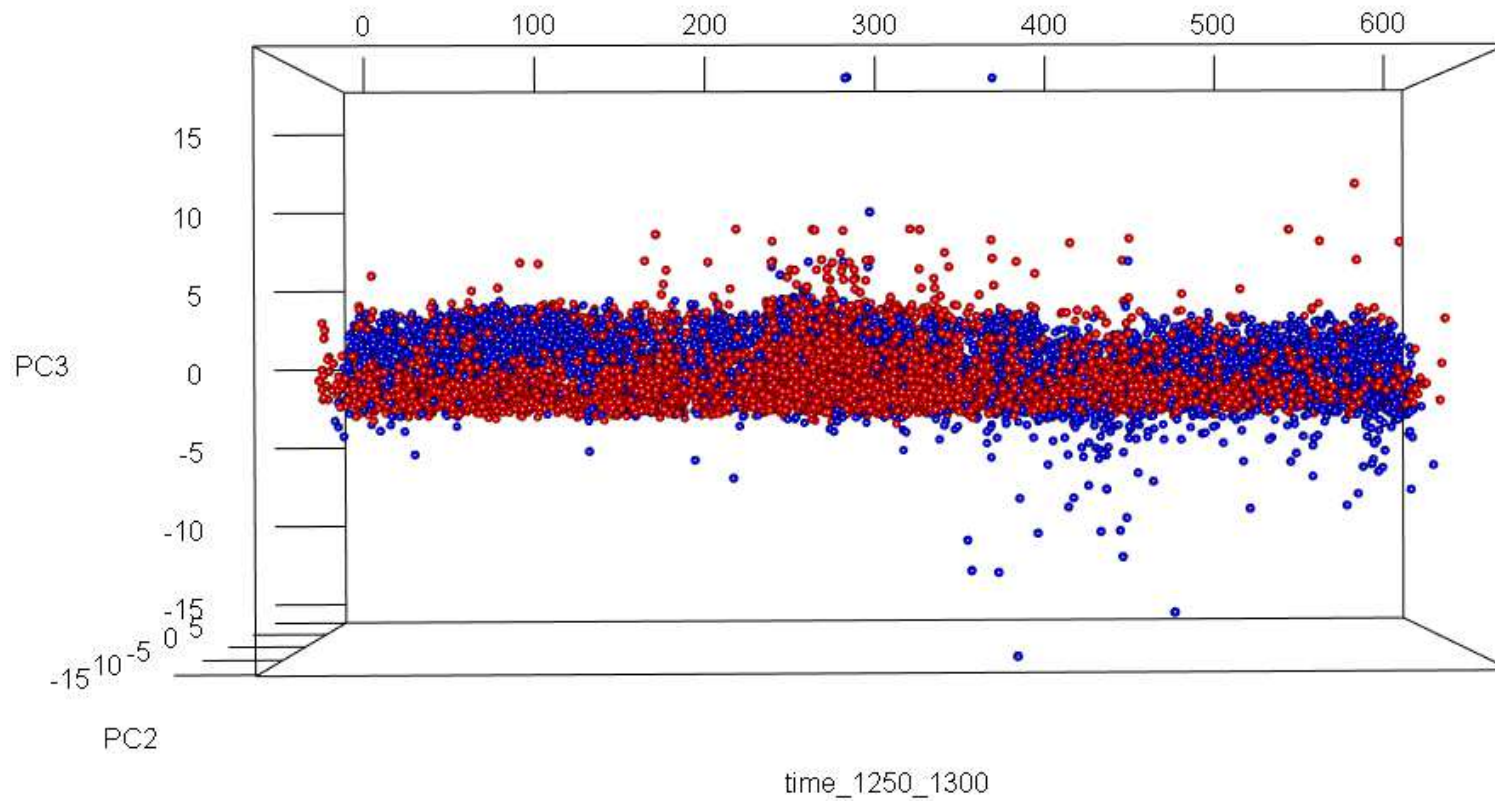


Figure 4.20 3D plots of PC2, PC3, and time in the dataset “unlabeled” with the view from PC3-time plane. The colors of points (blue and red) indicate the predicted class labels (snowflake and graupel), respectively.

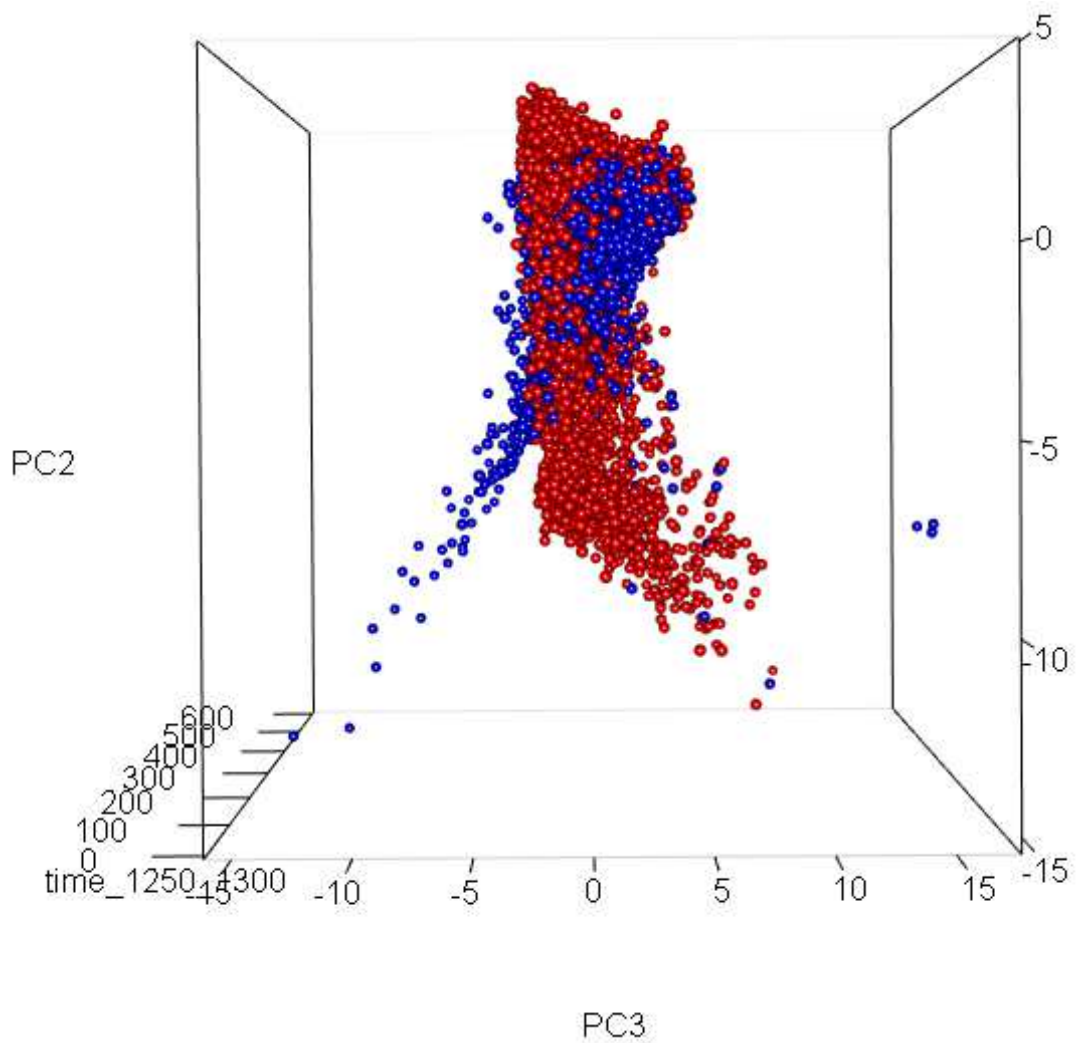


Figure 4.21 3D plots of PC1, PC2, and time in the dataset “unlabeled” with the view from PC1-time plane. The colors of points (blue and red) indicate the predicted class labels (snowflake and graupel), respectively.

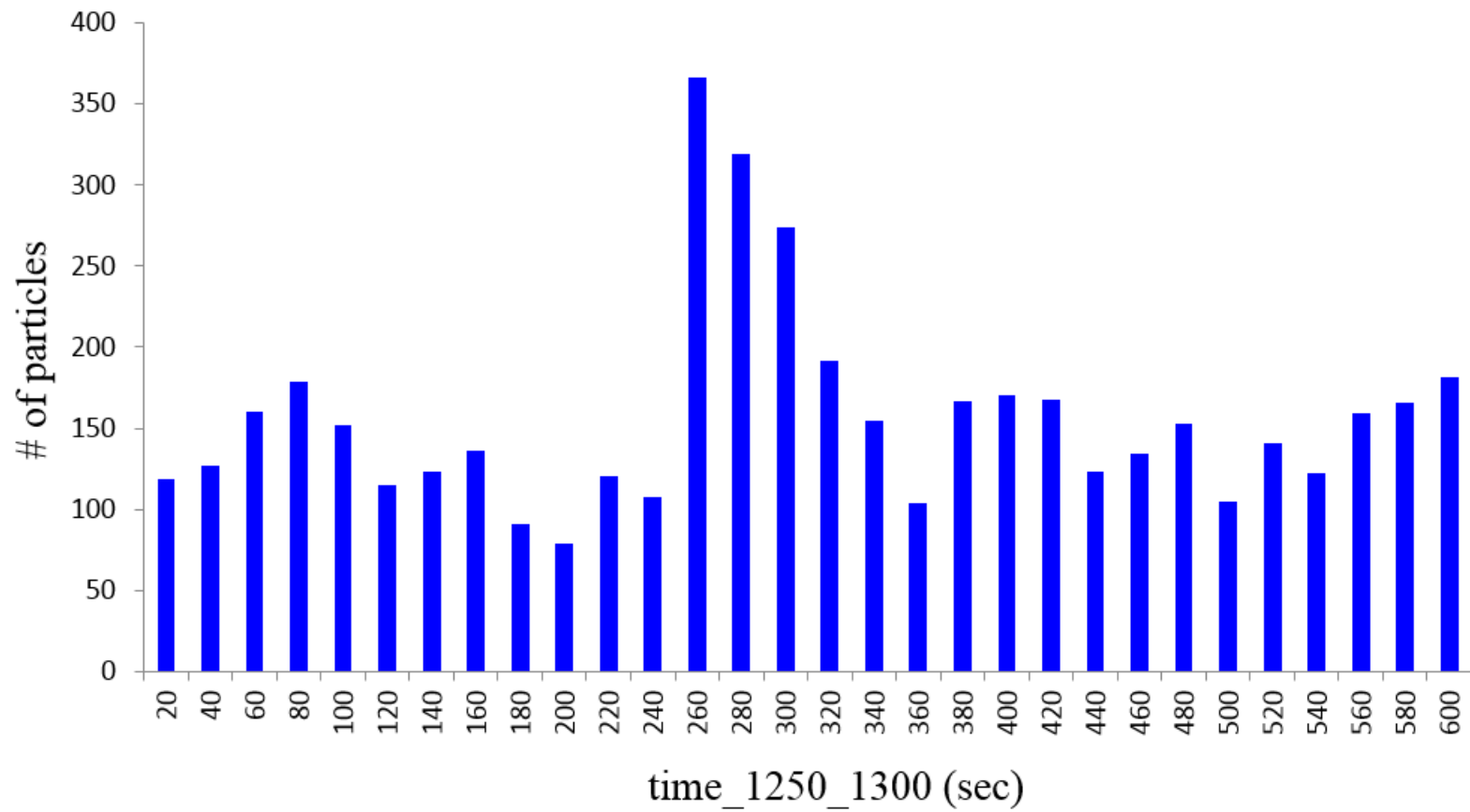


Figure 4.22 Histogram of predicted snowflakes in “unlabeled” dataset.

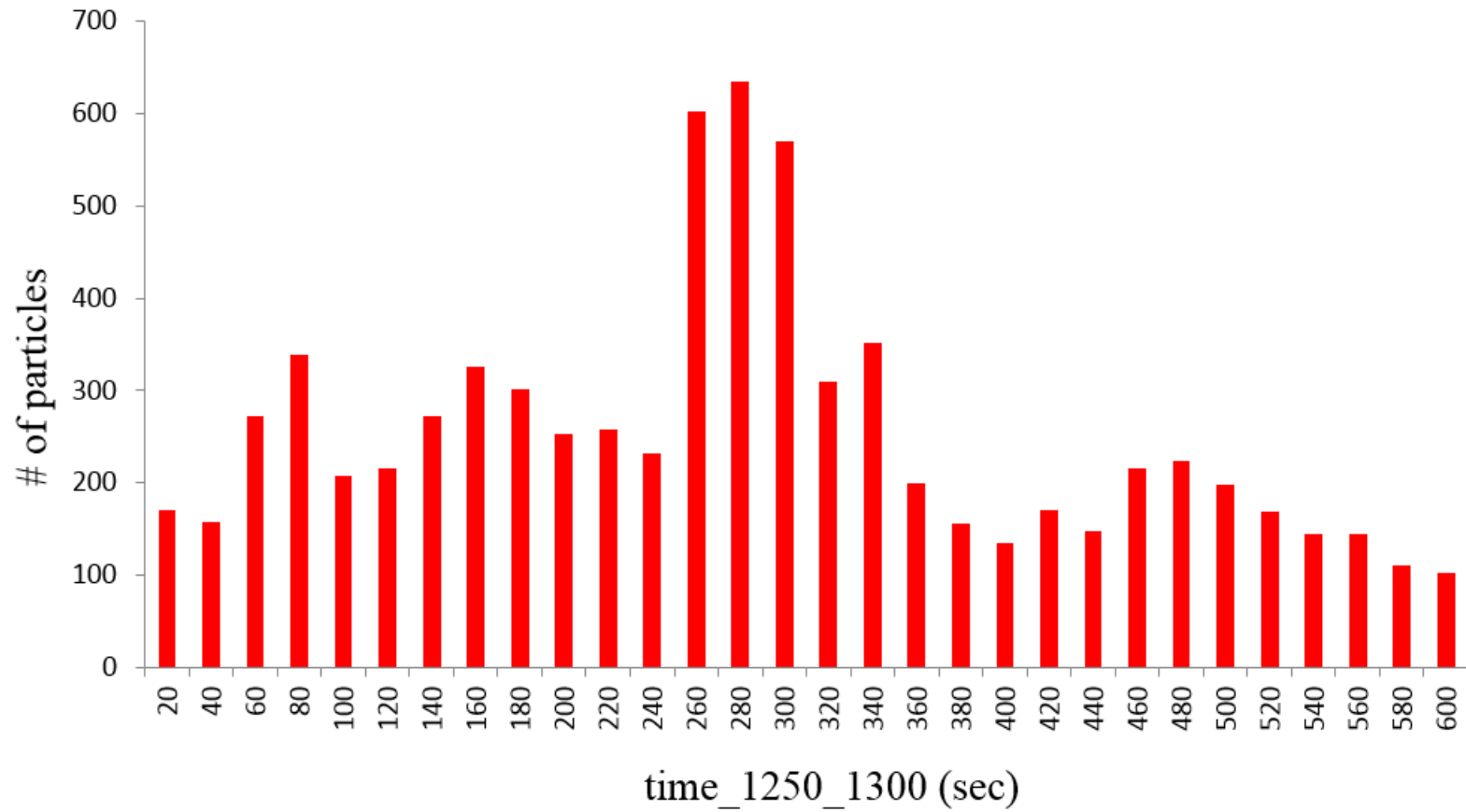


Figure 4.23 Histogram of predicted graupels in "unlabeled" dataset.

Chapter 5

Conclusion and Future Works

This final chapter is dedicated to review and summarize the accomplished work described in this paper. The closing section is also providing some ideas of future work vectors in the aim of improvement and modification.

5.1 Dissertation summary

Modern meteorological weather monitoring systems aim to enhance the knowledge in various research areas. One of them is snowfall formation mechanism. Proper on-ground observations are indisputably needed to justify the results of remote sensors work. One of the challenging tasks in this matter is to be able to perform the classification of solid precipitation into snowflake and graupel.

The ground-based measurements of snow particles and identification of snow type would be useful for deriving radar reflectivity-snow rate relationships.

In this study, we tried not only to (i) outperform the accuracy of the existing analogous classification methods, but to (ii) explicitly use the fractal features derived from particle shape and (iii) estimate the value of each feature in the contribution to classification. That was a nontrivial task due to the described study area problems. Moreover, it had been challenging as recent researches show significant advance in adjacent domains.

We conducted feature analysis and classification of particle data from 2DVD through the combined use of various statistical methods including supervised and unsupervised machine learning. Experimental results revealed that fractal and box-count features are useful for the characterization and classification of snowflakes

and graupel. The average accuracy of particle-by-particle classification was around 95.4%, which has not been achieved by previous studies.

From this result, it can be said that we could develop a system for automatically monitoring solid precipitation with practically sufficient accuracy of discriminating snowflakes and graupel. Additionally, we demonstrated that combining acquisition time information with the results of classification on large amount of particles, it becomes possible to conduct time-series analysis of amount and type of particles, which contributes to elucidate the mechanism of orographic snowfall (phenomena).

5.2 Future works

In this study, we mainly focused on two types of particles (i.e. snowflake and graupel). As an extension of this study, conducting human annotation with not only two types but also other detailed types of particles (e.g. dendrite-like, aggregate-like, melting-snow-like, and other depending on local precipitation particularity), makes it possible to quantitatively analyze wide-variety of snowfall in places with weather conditions not necessarily similar to those in Kanazawa. This may undoubtedly boost the practical applicability of the method yet lies beyond the scope of this study.

We hope that these two future work vectors will result in an even better method useful in a range of meteorological purposes.

Bibliography

- [1] Ohigashi, T. and Tsuboki, K. (2005) Structure and Maintenance Process of Stationary Double Snowbands along the Coastal Region. *Journal of the Meteorological Society of Japan*, **83**, 3, 331-349. <http://dx.doi.org/10.2151/jmsj.83.331>
- [2] Harimaya, T., Kodama, H. and Muramoto, K. (2004) Regional Differences in Snowflake Size Distributions. *Journal of the Meteorological Society of Japan*, **82**, 3, 895-903. <http://dx.doi.org/10.2151/jmsj.2004.895>
- [3] Brandes, E.A., Ikeda, K., Zhang, G., Schonhuber, M. and Rasmussen, R.M. (2007) A Statistical and Physical Description of Hydrometeor Distributions in Colorado Snowstorms Using a Video Disdrometer. *Journal of Applied Meteorology and Climatology*, **46**, 634-650. <http://dx.doi.org/10.1175/JAM2489.1>
- [4] Hung, G., Bringi, V.N., Cifelli, R., Hudak, R. and Petersen, W.A. (2010) A Methodology to Derive Radar Reflectivity–Liquid Equivalent Snow Rate Relations Using C-Band Radar and a 2D Video Disdrometer. *Journal of Atmospheric and Oceanic Technology*, **27**, 637–651. <http://dx.doi.org/10.1175/2009JTECHA1284.1>
- [5] Hung, G., Bringi, V.N., Moisseev, D., Petersen, W.A., Blivend, L., Hudake, D. (2014) Use of 2D-video disdrometer to derive mean density–size and Ze–SR relations: Four snow cases from the light precipitation validation experiment, *Atmospheric Research*, **153**, 34-48. <http://dx.doi.org/10.1016/j.atmosres.2014.07.013>
- [6] Zhang, G., Luchs, S., Ryzhkov, A., Xue, M., Ryzhkova, L. and Cao, Q. (2011) Winter Precipitation Microphysics Characterized by Polarimetric Radar and Video Disdrometer Observations in Central Oklahoma. *Journal of Applied Meteorology and Climatology*, **50**, 1558–1570. <http://dx.doi.org/10.1175/2011JAMC2343.1>
- [7] Locatelli, J.D., Hobbs, P.V. (1974) Fall speeds and masses of solid precipitation particles. *Geophys. Res.*, **79(15)**, 2185–2197 <http://dx.doi.org/10.1029/JC079i015p02185>

- [8] Michaelides, S.C., Levizzani, V., Anagnostou, E., Bauere, P., Kasparis, T., Lane, J.E. (2009) Precipitation: Measurement, remote sensing, climatology and modeling. *Atmospheric Research*, **94(4)**, 512–533
<http://dx.doi.org/10.1016/j.atmosres.2009.08.017>
- [9] Muramoto, K., Shiina, T., Endoh, T., Konishi, H., Kitano, K.,(1989) Measurement of snowflake size and falling velocity by image processing. *Proceedings of the NIPR Symposium on Polar Meteorology and Glaciology*, **2**, 48-54
<http://ci.nii.ac.jp/naid/110000029345>
- [10] Harimaya, T. (1988) The Relationship between Graupel Formation and Meteorological Conditions. *Journal of the Meteorological Society of Japan*. **66(4)**, 599-606
https://www.jstage.jst.go.jp/article/jmsj1965/66/4/66_4_599/_pdf
- [11] Hagen, M., Höller, H., Schmidt, K.(2012) Cloud and Precipitation Radar. *Atmospheric Physics Research Topics in Aerospace*, 347-361
http://dx.doi.org/10.1007/978-3-642-30183-4_21
- [12] Tanré, D., Artaxo, P., Yuter, S., Kaufman, Y. (2009) In Situ and Remote Sensing Techniques for Measuring Aerosols, Clouds and Precipitation. *Aerosol Pollution Impact on Precipitation*, 143-203
http://dx.doi.org/10.1007/978-1-4020-8690-8_5
- [13] DeWalle, D.R., Rango, A. (2008) Principles of Snow Hydrology. *Cambridge University Press*. <http://dx.doi.org/10.1017/CBO9780511535673>
- [14] Schuur, T.J., Ryzhkov, A. V., Zrnić, D. S. and Schönhuber, M. (2001) Drop Size Distributions Measured by a 2D Video Disdrometer: Comparison with Dual-Polarization Radar Data. *Journal of Applied Meteorology and Climatology*, **40**, 1019–1034.
[http://dx.doi.org/10.1175/1520-0450\(2001\)040<1019:DSDMBA>2.0.CO;2](http://dx.doi.org/10.1175/1520-0450(2001)040<1019:DSDMBA>2.0.CO;2)
- [15] Kruger, A. and Krajewski, W.F. (2002) Two-dimensional video disdrometer: a description. *Journal of Atmospheric and Oceanic Technology*, **19**, 602–617.
[http://dx.doi.org/10.1175/1520-0426\(2002\)019<0602:TDVDAD>2.0.CO;2](http://dx.doi.org/10.1175/1520-0426(2002)019<0602:TDVDAD>2.0.CO;2)

- [16] Nurzyska, K., Kubo, M. and Muramoto, K. (2010) 2D Feature Space for Snow Particle Classification into Snowflake and Graupel. *IEICE Transactions on Information and Systems*, E93-D, **12**, 3344-3351.
<http://dx.doi.org/10.1587/transinf.E93.D.3344>
- [17] Datla S, Sharma S (2008) Impact of cold and snow on temporal and spatial variations of highway traffic volumes. *Journal of Transport Geography***16**, 358–372 <http://dx.doi.org/10.4236/jts.2013.31003>
- [18] Changnon SA, Changnon D (2006) A spatial analysis of damaging snowstorms in the United States. *Natural Hazards***37**, 373–389
<http://dx.doi.org/10.1007/s11069-005-6581-4>
- [19] López-Moreno, J.I., Goyette S., Vicente-Serrano, S. M., Beniston, M. (2010) Effects of climate change on the intensity and frequency of heavy snowfall events in the Pyrenees. *Climatic Change*, **105**,489–508
<http://dx.doi.org/10.1007/s10584-010-9889-3>
- [20] Stoelinga, M.T., Hobbs, P. V., Mass, C. F., Locatelli, J. D., Colle, B. A., Houze JR, R. A., Rangno, A. L., Bond, N. A., Smull, B. F., Rasmussen, R. M., Thompson, G. and Colman, B. R. (2003) Improvement of Microphysical Parameterization through Observational Verification Experiment. *American Meteorological Society*, **105**,489–508
<http://dx.doi.org/10.1175/BAMS-84-12-1807>
- [21] Rasmussen, R., Baker, B., Kochendorfer, J., Meyers, T., Landolt, S., Fischer, A. P., Black, J., Thériault, J. M., Kucera, P., Gochis, D., Smith, C., Nitu, R., Hall, M., Ikeda, K. and Gutmann, E. (2012) How Well Are We Measuring Snow? *American Meteorological Society*, **93**, 811–829.
<http://dx.doi.org/10.1175/BAMS-D-11-00052.1>
- [22] Grazioli, J., Tuia, D., Monhart, S., Schneebeli, M., Raupach, T. and Berne, A. (2014) Hydrometeor classification from two-dimensional video disdrometer data. *Atmospheric Measurement Techniques*, **7**, 2869-2882.
<http://dx.doi.org/10.5194/amt-7-2869-2014>
- [23] Falconer, Kenneth, (2003) Fractal Geometry. *New York*. NY
<http://dx.doi.org/10.1002/0470013850>

- [24] Mandelbrot, B.B.(1982) *The Fractal Geometry of Nature*. W.H. Freeman and Company, San Francisco
- [25] Peitgen, H.-O., Saupe, D., Fisher, Y. and McGuire, M., (1988) *The Science of Fractal Images*. <http://dx.doi.org/10.1007/978-1-4612-3784-6>
- [26] Ishimoto, M. (2008) Radar Backscattering Computations for fractal-Shaped Snowflakes. *Journal of the Meteorological Society of Japan*, **86**, 3, 459-469. <http://dx.doi.org/10.2151/jmsj.86.459>
- [27] Maruyama, K. and Fujiyoshi, Y. (2005) Monte Carlo Simulation of the Formation of Snowflakes. *Journal of the Atmospheric Sciences*, **62**, 1529–1544. <http://dx.doi.org/10.1175/JAS3416.1>
- [28] Tolle, C.R., McJunkin, T.R., Gorsich, D.J. (2003) Suboptimal Minimum Cluster Volume Cover-Based Method for Measuring Fractal Dimension, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**, 1, 32-41. <http://dx.doi.org/10.1109/TPAMI.2003.1159944>
- [29] Russ, J.C. (2011) *The Image Processing Handbook*. Sixth Edition, CRC Press, 604-610. <http://www.crcpress.com/product/isbn/9781439840450>
- [30] Backes, A.R., Bruno, O.M. (2008) A New Approach to Estimate Fractal Dimension of Texture Images. *Image and Signal Processing*, 136-143 http://dx.doi.org/10.1007/978-3-540-69905-7_16
- [31] Jolliffe, I. (2005) Principal Component Analysis. *Encyclopedia of Statistics in Behavioral Science*. <http://dx.doi.org/10.1002/0470013192.bsa501>
- [32] Lu, D., Weng, Q. (2007) A Survey of Image Classification Methods and Techniques for Improving Classification Performance. *Int. J. Remote Sensing* **28(5)**, 823–870. <http://dx.doi.org/10.1080/01431160600746456>
- [33] Vapnik, V. (1998) *Statistical Learning Theory*, Wiley, New York, NY.
- [34] Sonka, M., Hlavac, V., Boyle, R. (2014) *Image Processing, Analysis, and Machine Vision*. Cengage Learning <http://compsadda.com/books/sem8/cvref.pdf>

[35] Fielding, A.H. (2006) Cluster and Classification Techniques for the Biosciences. *Cambridge University Press*.

<http://dx.doi.org/10.1017/CBO9780511607493>