Clustering analysis for classification and forecasting of solar irradiance in Durban, South Africa

by

Paulene Govender

Submitted in fulfilment of the academic requirements of

Doctor of Philosophy in Physics

School of Chemistry and Physics

College of Agriculture, Engineering and Science

University of KwaZulu-Natal, Westville

South Africa

December 2017

Abstract

Classification and forecasting of solar irradiance patterns has become increasingly important for operating and managing grid-connected solar power plants. A powerful approach for classification of irradiance patterns is by clustering of daily profiles, where a profile is defined as irradiance as a function of time. Classification is useful for forecasting because if the class of a day can be successfully forecast, then the irradiance profile of that day will share the general pattern of the class. In Durban, South Africa (29.871 °S; 30.977 °E), beam and diffuse irradiance profiles were recorded over a one-year period and normalized to a clear sky model to reduce the effect of seasonality, from which several variables were derived, namely minute-resolution beam, hourly-resolution beam and diffuse, and hourly-resolution beam variability. To these variables, individually and in combination, *k*-means clustering was applied, and beam irradiance was found to be the one that best distinguishes between sky conditions. In particular, clustering of hourly-resolution beam irradiance produced four classes with diurnal patterns characterized as sunny all day, cloudy all day, sunny morning-cloudy afternoon, and cloudy morning-sunny afternoon. These classes were then used to forecast beam and diffuse irradiance for the day ahead, in association with cloud cover forecasts from Numerical Weather Prediction (NWP) output. Two forecasting methods were investigated. The first used *k*-means clustering on predicted daily cloud cover percentage profiles from the NWP, which was a novel aspect of this research. The second used a rule whereby predicted cloud cover profiles were classified according to whether their averages in the morning and afternoon were above or below 50%. From both methods, four classes were obtained that had diurnal patterns associated with the irradiance classes, and these were used to forecast the irradiance class for the day ahead. The two methods had a comparable success rate of about 65%. In addition, hour-ahead forecasts of beam and diffuse irradiance were performed by using the mean profile of the forecast irradiance class to extrapolate from the current measured value to the next hour. The method showed an average improvement of about 22% for beam and diffuse irradiance over persistence forecasts. These results suggest that classification of predicted cloud cover and irradiance profiles are potentially useful for development of class-specific, multi-hour irradiance forecast models.

Preface

The research contained in this thesis was carried out in the School of Chemistry and Physics, University of KwaZulu-Natal, Westville, Durban, from June 2014 to December 2017, under the supervision of Dr A. P. Matthews and co-supervision of Dr M. J. Brooks. The research was financially supported by the National Research Foundation (NRF).

These studies represent original work by the author and have not otherwise been submitted in any form for any degree or diploma to any tertiary institution. Where use has been made of the work of others it is duly acknowledged in the text.

As the candidate's supervisor I have approved this dissertation for submission.

Signed: _____ Name: _____ Date: _____

Declaration 1-Plagiarism

I, Paulene Govender declare that

1. The research reported in this thesis, except where otherwise indicated or acknowledged, is my original work.

2. This thesis has not been submitted in full or in part for any degree or examination to any other university.

3. This thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.

4. This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:
   (a) their words have been re-written but the general information attributed to them has been referenced.
   (b) where their exact words have been used, their writing has been placed inside quotation marks, and referenced.

5. This thesis is primarily a collection of material, prepared by myself, published as journal articles or presented as a poster and oral presentations at conferences. In some cases, additional material has been included.

6. This thesis does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the dissertation and in the Bibliography section.

Signed: .....

Declaration 2-Publications

1. **Govender P.**, Brooks M. J. and Matthews A. P. (*in press*). "Cluster analysis for classification and forecasting of solar irradiance in Durban, South Africa". Journal of Energy in Southern Africa.

2. Bessafi M., Delage O., Jeanty P., Heintz A., Cazal J. D., Delsaut M., Gangat Y., Partal L., Lan-Sun-Luk J. D., Chabriat J. P., Brooks M. J., Matthews A. P., Pitot J. and **Govender P.** (2015). "Solar research collaboration between University of KwaZulu-Natal and University of Reunion Island". Proceedings of the 3^{rd} Southern African Solar Energy Conference (SASEC), May 2015, Kruger National Park, Skukuza, South Africa.

3. Matthews A. P., **Govender P.**, Ganya E., Brooks M. J., Venkataraman S., "Solar radiometry and forecasting research at UKZN". Proceedings of the 60^{th} Annual Conference of the South African Institute of Physics (SAIP), July 2015, Nelson Mandela Metropolitan University, Port Elizabeth, South Africa.

For my parents.

Acknowledgements

- Firstly, I would like to express my sincere gratitude to my supervisor Dr Alan Matthews for his contribution towards the completion of this work. Your invaluable guidance, motivation, patience and insightful discussions throughout this research is highly appreciated.

- To my co-supervisor, Dr Michael Brooks, thank you for all your constructive criticism and advice during the course of this work. I really appreciate all of your guidance and assistance especially with the solar instrumentation.

- I would also like to acknowledge our collaborators from the University of La Réunion, Professor Miloud Bessafi and Mr Mathieu Delsaut. Thanks to Professor Miloud Bessafi for hosting me at the Electronic, Energy and Processes Laboratory (LE^2P). I would like to especially thank Mr Mathieu Delsaut for his helpful discussions and for sharing with me his vast knowledge on clustering techniques during my time in Réunion.

- I would also like to thank Evans Zhandire for his assistance with MATLAB software whenever it was needed.

- To my parents, thank you for all the encouragement and unwavering support that you have always given me.

- To Rakesh Mohanlal, thank you for your never ending motivation and inspiration, and for sharing with me your own PhD experience.

- Lastly, I would like to thank the National Research Foundation (NRF) for financial support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter introduces solar forecasting and its importance for solar power plants. It briefly describes solar forecasting methods and some previous studies that used them. The chapter sets out the different approaches that were considered during the development of a forecasting model for Durban, and the reasons that led to the use of clustering and cloud cover forecasts, to produce day-ahead irradiance forecasts. It concludes with a list of research objectives and a thesis outline.

1.1 Solar forecasts for power plants

Forecasting of solar energy has become an important tool for efficiently operating and managing grid-connected solar energy plants. The variability in irradiance at ground level results in fluctuation of power output of solar power plants, causing grid instability and uncertainty in power output. With the growing number of solar power plants, the need for forecasting technologies is increasing as they become invaluable for grid operators who manage installations.

The variable nature of solar energy at ground level is due to clouds, aerosols and water vapour. Of these, clouds are dominant and therefore there is a need to predict their amount, velocity and transmissivity (Marquez and Coimbra, 2013). Cloud movement, having a highly stochastic nature, presents a significant challenge to achieving an accurate depiction of the local cloud distribution. Cloud properties such as size and spatial distribution, composition (vapour, liquid or ice particles) and opacity are some of the characteristics responsible for their complex interaction with radiation. These factors significantly contribute to the high level of difficulty associated with the development of efficient solar forecasting models.

With the addition of large-scale solar power plants i.e. photovoltaic (PV) and concentrated solar thermal (CST) systems, grid-related activities such as load following and management, power scheduling and unit commitment, and maintenance scheduling as well as other necessary plant operations will greatly benefit from solar forecasts (Inman et al., 2013). Knowledge of irradiance levels will enable plant operators to efficiently manage the above-mentioned activities and lower operational and maintenance costs by limiting the use of ancillary devices such as back-up generators.

1.2 Classification of irradiance

The development of an effective forecasting model is contingent on a proper understanding of solar irradiance patterns at a location of interest. This is key to understanding and characterizing the solar resource at a location. A powerful approach to understanding solar irradiance patterns is by classification and characterization of irradiance profiles using cluster analysis (or clustering). A profile is defined as irradiance as a function of time over a day. An example of a profile is given in Figure 1.1, where the profile is the minute-resolution direct (beam) irradiance component and is denoted as B. This thesis investigates the use of clustering for classification and forecasting of irradiance profiles in Durban. Located on the east coast of South Africa, Durban is a region with humid sub-tropical climate and significant cloud variation. Although its direct normal, global and diffuse irradiance characteristics have been described (Lysko, 2006; Zawilska and Brooks, 2011), limited work has been done to characterize the irradiance patterns using a clustering approach. Furthermore, for Durban, there has been limited work related to classification of irradiance profiles for forecasting.

As seen from the solar forecasting literature (Diagne et al., 2013; Inman et al., 2013; Kleissl, 2013), some well-known and commonly-used forecasting methods include statistical (linear and non-linear), image-based (using satellite and ground-based sensors) and Numerical Weather Prediction (NWP) methods. Statistical methods are applied to time series data and involve analyzing their past patterns to make a forecast. These include techniques such as Linear Regression (LR), Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA) and Artificial Neural Networks (ANNs). Image-based methods use cloud imagery, either from a satellite or a ground-based device, to track cloud motion. This is achieved by applying cloud motion vectors (CMVs) to consecutive cloud images and, based on their speed and trajectory, future

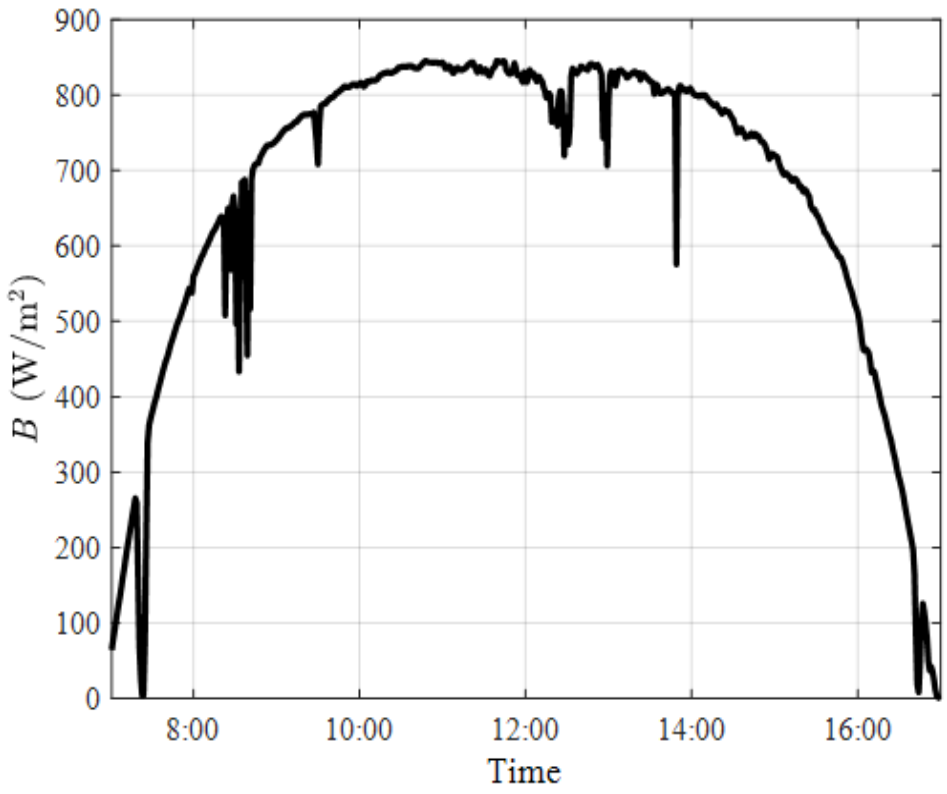


Figure 1.1: Typical irradiance profile for the minute-resolution beam irradiance component, B, on a mostly clear day for 9 June 2017 in Durban.

cloud positions are predicted. The NWP method uses mathematical models to simulate the state of the atmosphere based on a set of initial conditions. Atmospheric variables such as temperature, precipitation, humidity, pressure, wind speed and direction and cloud cover are available as outputs from NWP forecasts (Chaturvedi, 2016; Kleissl, 2013). Of particular interest in the present work is the cloud cover forecast from the NWP.

Clustering of daily irradiance profiles forms the basis of this thesis, and the main aim was to investigate the use of clustering for forecasting. In general, the aim of cluster analysis is to identify groups of similar objects, where objects in a cluster are more similar to each other than objects in different clusters (Halkidi et al., 2001). Clustering therefore can reveal information about the data that may have been previously unnoticed. For the present work, clustering of daily irradiance profiles was used to conduct a classification and characterization of the solar irradiance patterns in Durban. Classification is useful for forecasting because if the class of a day can be successfully forecast, then the irradiance profile of that day will share the general pattern of the class. For the

classification and characterization of solar irradiance patterns in Durban the first part of this thesis, and the first objective of this research, presents a clustering of minute-resolution beam irradiance profiles.

Several early studies, for example Zangvil and Lamb (1997), Muselli et al. (2000), Muselli et al. (2001), Maafi and Harrouni (2003), Diabaté et al. (2004), Harrouni et al. (2005), Soubdhan et al. (2009), Gastón-Romeo et al. (2011) and Kang and Tam (2013) have investigated classification of days based on solar irradiance profiles. The most commonly-used parameter among these studies was the clearness index, k_t, which is the ratio of the measured global horizontal irradiance (GHI) at the Earth's surface to the Top of Atmosphere (TOA) or extraterrestrial irradiance. In addition, these studies focused mainly on classification of irradiance profiles and did not consider forecasting. Studies that are more closely related to the present work include those of Badosa et al. (2013), Badosa et al. (2015), Jeanty et al. (2013), McCandless et al. (2014) and McCandless et al. (2015). McCandless et al. (2014) used clustering applied to k_t to identify cloud regimes and thereafter solar irradiance models were developed specifically for each regime. In McCandless et al. (2015), 7 cloud regimes were identified and used for forecasting, and several NWP outputs including cloud cover were considered, but were not used for forecasting. Badosa et al. (2015) explored the use of exogenous variables such as synoptic wind and humidity for day-ahead irradiance forecasts. Although cloud cover was not used, the novelty of the study lay in using only exogenous variables in the model. The method was successfully used for day-ahead irradiance forecasts. Similar to the present work, Jeanty et al. (2013) used clustering for classification of the irradiance profiles in Reunion Island, and the establishment of classes that describe the diurnal patterns. Even though the authors indicate that the classes can be used for forecasting, the study did not pursue this avenue of investigation. Instead, through clustering of a single variable, the study was limited to providing a classification and characterization of the solar irradiance patterns in Reunion Island. The studies discussed above provide a basis where much of the present study focuses on clustering and classification, and how it can be applied to forecasting.

1.3 Development of a solar forecasting model for Durban

In the development of a forecasting model for Durban, different forecasting methods were investigated including LR, ARIMA and ANNs. LR and ARIMA use a fit to time series data to extrapolate

and predict future values. Examples of such studies that have used these approaches include Paoli et al. (2010), Voyant et al. (2011), Voyant et al. (2012), Paoli et al. (2014) and Lauret et al. (2016). A limitation of these time series approaches is that they are unable to accurately predict deviations from the irradiance trend that are caused by clouds, and are referred to as excursions. With reference specifically to beam irradiance, excursions may occur during clear days where the irradiance decreases from clear sky values to close to zero due to the presence of clouds obscuring the Sun. Alternatively, excursions may occur during cloudy days where the irradiance increases from zero to almost clear sky values due to gaps in the clouds.

Excursions are difficult to predict using only irradiance time series and it was for this reason that LR and ARIMA techniques were not used in the present work. ANNs, on the other hand, are able to model more complex behaviour in time series data but require a large amount of data for training of the network. Due to the lack of uninterrupted, long-term meteorological and irradiance records in Durban, ANNs were also not used. Furthermore, in an attempt to combine different data sources in the forecasting method, the use of satellite imagery was also investigated. However, due the high cost associated with acquiring high spatial and temporal resolution imagery over Durban, this avenue was not pursued. Considering these factors, investigating the use of clustering and classification of available radiometric data for forecasting emerged as a promising avenue.

For the classification and characterization of the irradiance patterns in Durban, clustering was applied to minute-resolution profiles of the beam irradiance, normalized to a clear sky model, over a period of one year. As will be shown in Chapter 5, the horizontal beam irradiance variable used for clustering by Jeanty et al. (2013) is dependent on seasonality. Although the present work is similar to that of Jeanty et al. (2013), clustering was applied to the normalized beam irradiance. The normalized beam irradiance removes seasonal dependency and hence any variation in the signal is mainly due to cloud conditions. Since this study investigates clustering of irradiance profiles under different cloud conditions, introducing the normalized beam irradiance was appropriate.

Due to the high number of dimensions present in the minute-resolution normalized beam irradiance profiles, a pre-processing technique was applied prior to clustering. The pre-processing technique was Principal Component Analysis (PCA), that takes high-dimensional data and reduces it to a lower dimension while retaining most of the information. As described by Jolliffe (2002), this is achieved by transforming to a new set of axes, called Principal Components, which are ordered successively so that the first few components retain most of the variance present in the original data.

The beam and corresponding diffuse irradiance classes resulting from the pre-processing and clustering, characterize the diurnal irradiance patterns in Durban.

As mentioned earlier, cloud cover forecasts are of particular interest for the present work. Cloud cover forecasts from NWP output are publicly-available from a weather-service provider at hourly-resolution for the day ahead. One of the aims of the present work was to investigate how clustering of irradiance and the resulting classes can be combined with cloud cover forecasts from the NWP, to forecast irradiance for the day ahead. One of the steps to achieving this was to apply clustering to cloud cover profiles, and to correlate them with irradiance clusters. However, since cloud cover profiles were only available at hourly-resolution, the original minute-resolution irradiance profiles were reduced to hourly-resolution to match the temporal resolution of the cloud cover forecasts. Clustering was then applied to the hourly-resolution irradiance profiles and it will be shown that (1) hourly-resolution profiles produce the same clustering as minute-resolution profiles and (2) the clustering of the hourly-resolution beam irradiance profiles produces classes that can be associated with those from the clustering of cloud cover. The clustering of cloud cover output from the NWP, and its correspondence with clustering of beam irradiance for day-ahead forecasting, is a novel aspect of this thesis.

The conversion of beam irradiance profiles from minute-resolution to hourly-resolution can result in a loss of information. Therefore in order to regain some information that was originally contained in the minute-resolution data, variability in the beam irradiance profiles was also investigated. Variability that in this context is the short-term fluctuations in the irradiance due to clouds that occur at the minute-timescale. Studies that have applied clustering to irradiance variability include that of Watanabe et al. (2016) and Zagouras et al. (2013). However, these studies focused on spatial variability over different geographical regions, rather than temporal variability at hourly-scale for a single geographical location, as in the present work. For the present work, variability in the beam irradiance derived from the minute-resolution data, and clustering were applied to the resulting hourly-resolution beam variability profiles. In addition, investigating the beam variability profiles led to clustering the combination of beam irradiance and its variability, to form a two-variable clustering set. This combination was intended to investigate whether a stronger clustering solution emerges to forecast irradiance for the day ahead.

For this thesis, forecasting daily irradiance class profiles uses cloud cover forecasts from the NWP in combination with clustering of irradiance profiles. For day-ahead irradiance forecasts, two

methods were investigated. The first uses clustering of cloud cover and the second uses a simple set of decision rules called the "50% rule". The methods provide a day-ahead forecast of the class mean profiles for both beam and diffuse irradiance quantities. The reason for forecasting beam and diffuse quantities is that they are both important for solar power plants. Furthermore, global irradiance can be found if both quantities are known.

In addition to forecasting the irradiance class for the day-ahead, forecasts for individual hourly values of beam and diffuse irradiance i.e. hour-ahead forecasts are presented. The method used is called "Persistence of the Class Trend". The method also makes use of the NWP cloud cover forecast to classify the day into a class. Thereafter, the method uses the actual measured value and the class mean for the current hour, and the class mean for the next hour to predict beam and diffuse irradiance for the hour ahead. It will be shown that in some cases the method is an improvement from traditional Persistence because it uses the class mean profile as a reference against which the prediction for the next hour can be adjusted.

In summary, this investigation focuses on the use of clustering and classification of irradiance combined with cloud cover forecasts for day-ahead irradiance forecasts. The clustering of several irradiance variables which could be used for forecasting was investigated. In addition to minute-resolution normalized hourly-resolution beam, hourly-resolution diffuse irradiance and variability in the normalized beam irradiance were also clustered. To investigate which would be most useful for forecasting, some of these variables were clustered separately and some combined to form a multi-variable clustering set. From the clustering of the several irradiance variables, it is shown that beam irradiance gives the best clustering solution for combination with cloud cover, for forecasting irradiance in Durban.

1.4 Research objectives

The specific objectives of this research are as follows:

- Use clustering analysis techniques to classify and characterize solar irradiance patterns in Durban, South Africa.
- Investigate several irradiance clustering variables, for example normalized beam irradiance and the combination of the normalized beam irradiance with its variability, to find which gives the best clustering solution that can be used for day-ahead forecasting.
- Combine Numerical Weather Prediction cloud cover forecasts and the clustering results to develop a model for day-ahead forecasting of daily normalized beam and diffuse irradiance profiles for Durban.

1.5 Thesis outline

This thesis is structured into seven chapters. Chapter 2 contains a general overview of terrestrial solar radiation. Ground-based solar radiation instrumentation for measurement of the three main irradiance components is described. The normalization of the irradiance components was done by the use of a clear sky model, whose selection and implementation are highlighted. The variables used for clustering in the remaining chapters are defined.

Chapter 3 gives an overview of the different forecasting methods with their relevant strengths and weaknesses. Three forecasting horizons are defined and a review of the solar forecasting literature within each of the forecasting horizons is presented. This chapter presents an illustration that summarizes the most effective forecasting method for the different forecasting time horizons.

Chapter 4 begins with an introduction of cluster analysis and reviews the literature on the application of clustering to solar irradiance measurements. The theory of dimension reduction, a pre-processing step to cluster analysis and cluster analysis techniques which form the basis for the development of a forecasting model for this research, are presented. In addition, a description of how dimension reduction and clustering techniques are applied to understand, characterize and classify solar irradiance patterns in Durban is given. A set of minute-resolution horizontal beam irradiance profiles are used to illustrate the dimension reduction and clustering techniques.

Chapter 5 details the classification of irradiance profiles using clustering. Clustering is applied to several variables which are derived from radiometric data. Some variables are clustered on their own and some in combination. Initial results and comparisons for each of the clustered variables are presented. From clustering the normalized beam irradiance profiles, a set of classes that classify and characterize the solar irradiance patterns in Durban are established.

Forecasting using the classes established in Chapter 5 is the focus of Chapter 6. First, day forecasts (in the form of a class) are produced using two methods and second, hourly forecasts for individual values of beam and diffuse irradiance are presented. Details of each of the forecasting methods are described. Results such as the success rate of the forecasting methods, in the case of the day forecasts, and quantification of the forecast error using an appropriate error metric, in the case of hourly forecasts, are given.

A discussion of the clustering and forecasting results from previous chapters is presented in Chapter 7. This includes a discussion of the clustering solution that provides the best characterization of the solar irradiance patterns in Durban and a comparison between forecasting methods. Chapter 8 gives a conclusion of the main findings of this study.

Chapter 2

Solar radiation and instrumentation

This chapter includes an overview of solar radiation at ground level and its variation throughout the year. Measurement of the solar resource using ground-based instrumentation is discussed, and details of the sensors used for measuring direct (beam), diffuse and global irradiance are given. The selection of a clear sky model is motivated and examples of clear sky profiles for Durban are presented. The chapter concludes with the definition of a set of radiometric variables that are used for clustering, and examples of the normalized daily beam and diffuse irradiance profiles.

2.1 The solar spectrum

The radiation emitted from the Sun is in the form of electromagnetic waves of varying wavelength, λ. Most of the electromagnetic energy is concentrated in the ultraviolet ($100 < \lambda < 400$ nm), visible ($400 < \lambda < 700$ nm) and infrared ($700 < \lambda < 1 \times 10^6$ nm) regions of the spectrum (Goswami et al., 1999). Some of these wavelengths are screened from the Earth's surface by different layers of the atmosphere. The main wavelength reaching the surface is that of visible light. However, portions on either side of the visible spectrum are not completely cut off and are able to pass through the atmosphere. These are the ultra-violet and infrared ranges. The infrared range has wavelengths that are too long to be seen by the naked eye, whereas the ultra-violet region contains wavelengths that are too short. The solar spectrum is approximately equal to that of a black body at a temperature of 5800 K. Due to the absorption of certain frequencies by atmospheric constituents such as water vapour, dust particles, ozone and other molecules in the air, the spectrum that is received by the Earth's surface is significantly altered (Sen, 2008). Figure 2.1 shows the solar radiation spectrum.

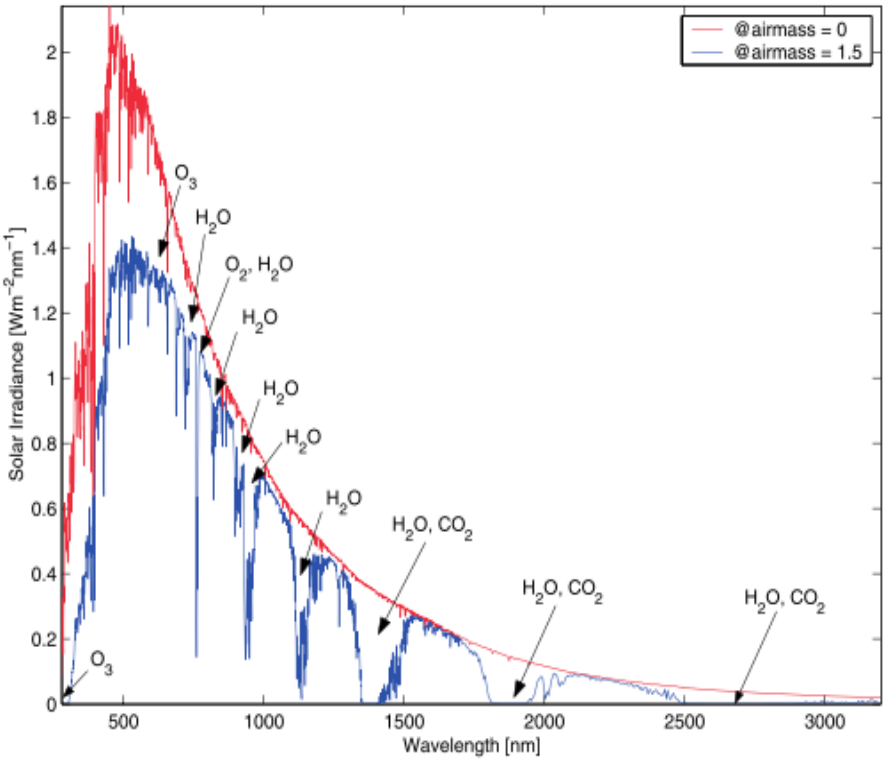


Figure 2.1: Spectral distribution of solar energy above the atmosphere (red) at zero air mass and the altered spectrum (blue) that reaches the surface of the Earth at air mass equal to 1.5. Adapted from Lysko (2006).

As described by Duffie and Beckman (1991), *irradiance* is the rate at which radiant energy is incident on a surface, per unit area of surface. It is typically measured in W/m^2. The amount of solar radiation reaching the surface of the Earth is greatly reduced by the absorption, reflection and scattering of light. Factors such as air molecules, clouds, dust particles and water vapour contribute to the reduction of the amount of sunlight received. Radiation that is scattered is referred to as *diffuse* solar radiation. The radiation that does not undergo scattering, absorption or reflection is known as *direct* or *beam* radiation. The direct and diffuse solar radiation can be summed to give the *global* solar radiation. According to Myers (2005), the global (sky+solar disk) radiation, G, incident on a horizontal surface is the sum of the beam radiation, B, projected onto the surface and hence modified by the cosine of the incidence angle of the beam, i_b, and diffuse sky radiation, D, from the dome of the sky excluding the Sun. This relationship is expressed by

$$G = B\cos(i_b) + D. \tag{2.1}$$

2.2 Sun-Earth geometry

The orbit of the Earth around the Sun is slightly elliptical with an eccentricity ϵ (i.e. the ratio of major to minor axis) of 0.0167. The eccentricity of the orbit produces changes in the Earth-Sun distance, or "radius vector". The speed of the Earth varies within its elliptical orbit i.e. is higher at the perihelion (the closest approach of the Earth to the Sun) than at the aphelion (the farthest approach of the Earth to the Sun) (Myers, 2013).

The obliquity of the ecliptic i.e. the plane of the Earth's annual orbital motion around the Sun is another factor that contributes to variation in the Earth's motion. As stated in Muller (1995), the equal angles which the Sun in its apparent movement goes through in the ecliptic, does not correspond to equal angles that are measured on the equatorial plane. The angles measured on the equatorial plane are relevant for the measurement of time, since the daily movement of the Sun is parallel to the equatorial plane. The angle between the equatorial plane and the ecliptic is the declination angle, as will be defined in equation (2.2). The declination can be regarded as the deviation between the projected equatorial plane and the orbital plane. The ecliptic is inclined to the celestial equator by an amount equal to the declination, and it is this tilt that gives rise to the obliquity. Therefore, the eccentricity of the Earth's orbit and the obliquity of the ecliptic are the two factors that are responsible for the Equation of Time (EOT).

The amount of irradiance received at ground level varies with the Earth's tilt and its motion as it orbits the Sun. Due to the eccentricity of the Earth's orbit, there is a 3% increase in the solar radiation intensity at perihelion and a 3% decrease at aphelion. The 23.45 $^\circ$ axial tilt of Earth's rotation axis, with respect to the plane of Earth's orbit, produces the seasonal weather changes (Myers, 2013). This angle is termed the *declination* angle and denoted as δ. The declination angle also causes the daily variation in the points on the horizon where the sun rises and sets, the path of the sun through the sky, and the day length Myers (2013). Therefore, depending on the season, the amount of irradiance received at the surface of the Earth varies. The elliptical orbit of the Earth's path around the Sun is shown in Figure 2.2. It indicates the position of the Earth relative to the Sun at different times of the year i.e. during the equinoxes and the solstices. For the southern hemisphere the equinoxes occur in March and September and the winter and summer solstices in June and December, respectively.

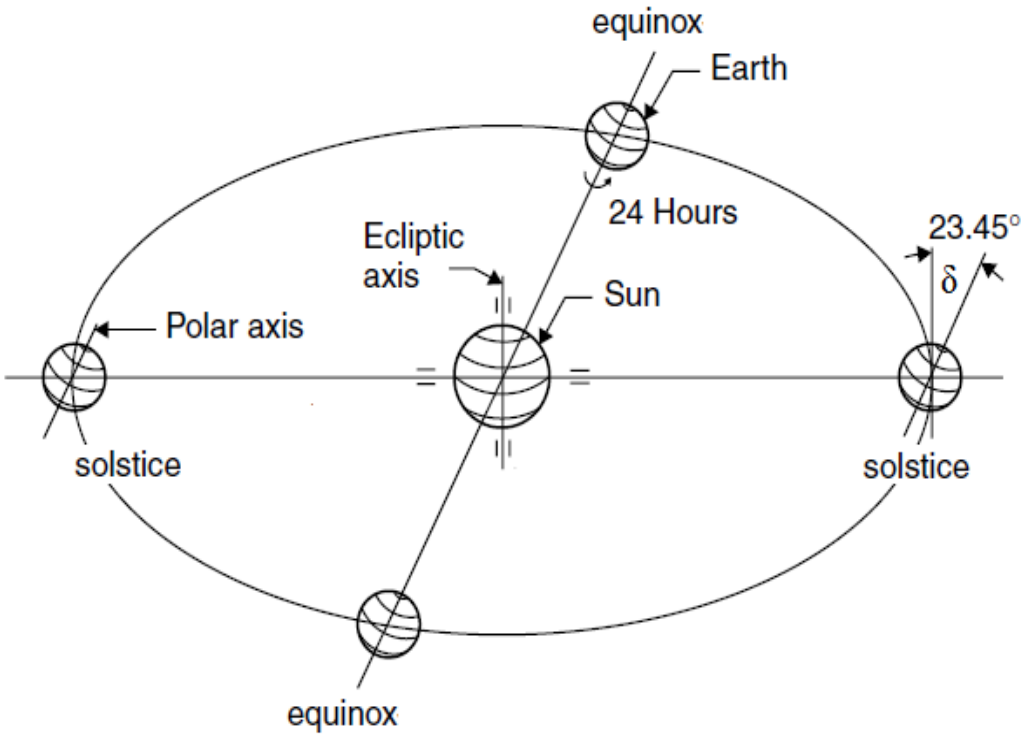


Figure 2.2: Motion of the Earth about the Sun. For the southern hemisphere the equinoxes occur in March and September and the winter and summer solstices in June and December, respectively. Adapted from Kalogirou (2009).

The declination angle varies through the year and is at its extreme during the solstices (longest or shortest length of day) and at zero during the equinoxes (equal day and night). The variation of this angle, is shown in Figure 2.3 and the declination angle can be approximated by

$$\delta = \delta_o \left(\sin \frac{360\,^\circ (284 + n)}{365} \right), \tag{2.2}$$

where δ varies between $+\delta_o$ = $+23.45\,^\circ$ (mid-summer in the northern hemisphere) and $-\delta_o$ = $-23.45\,^\circ$ (mid-winter in the northern hemisphere) and *n* = day number. (Twidell and Weir, 2006).

The time it takes Earth to traverse equal distances along its elliptical orbit varies through the year, while the Earth's daily rotation rate is constant. Therefore, the local standard time that the Sun is located on the local meridian (local solar noon) will vary through the year. For solar energy calculations, apparent solar time (ST) is used to express the time of day. Apparent solar time is

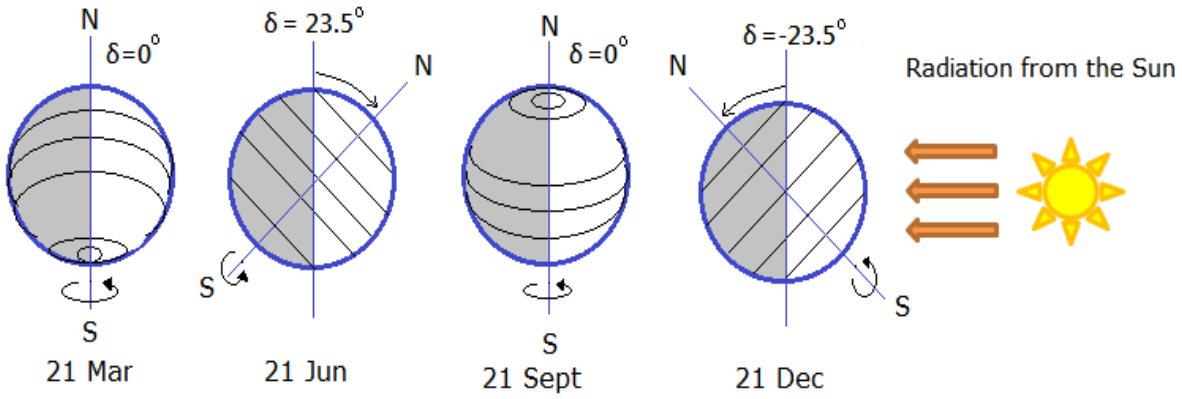


Figure 2.3: Variation of the declination angle through the year. The Earth's circles of latitude are shown. A declination angle of 0 ° denotes the two equinoxes. Adapted from Twidell and Weir (2006).

based on the apparent angular motion of the sun across the sky and does not coincide with clock time (Duffie and Beckman, 1991). The difference between standard time and solar time is the *EOT*.

The relationship between apparent solar time and the local time at a location, referred to as the local standard time (LST), is calculated using the following equation:

$$ST = LST + EOT + (l_{st} - l_{local}) \cdot 4min/degree, \tag{2.3}$$

where l_{st} and l_{local} is the standard meridian of the local time zone and local longitude, respectively and *EOT* is time in minutes approximately given by

$$EOT = 9.87\sin(2B_{EOT}) - 7.53\cos(B_{EOT}) - 1.5\sin(B_{EOT}), \tag{2.4}$$

where $B_{EOT} = 360(n-1)/364$ degrees (Goswami et al., 1999).

The variation in the intensity of irradiance reaching the surface of Earth induced by the rotation of the Earth and its position relative to the Sun, is predictable. This variation is not noticeable for very short time intervals i.e. seconds to minutes, but becomes prominent for longer time intervals (Kleissl, 2013). The short-term variation in irradiance that is less predictable, and that is of relevance to the operation of solar power plants, is due to clouds.

2.3 Interaction of clouds with irradiance

A cloud is described by Lohmann et al. (2016) as "an aggregate of water droplets or ice crystals, or a combination of both, suspended in air". As described by Rogers and Yau (1989), the existence of

clouds is due to the physical process of condensation. Condensation occurs primarily in response to dynamic processes that include widely distributed vertical air movement, convection and mixing. Within a cloud system there can be individual cloud structures or elements of a cloud that can be identified by their shape and size. Their sizes could range from 1 to 100 km in spatial extent and have lifetimes from minutes to hours. Together, they constitute the cloud system. As described by Tapakis and Charalambides (2013), clouds can also be classified based on their altitude:

- Low level clouds: Cumulus (Cu), Stratocumulus (Sc), Stratus (St), and Cumulonimbus (Cb).
- Mid-level clouds: Altocumulus (Ac), Altostratus (As) and Nimbostratus (Ns).
- High level clouds: Cirrus (Ci), Cirrocumulus (Cc) and Cirrostratus (Cs).

Their composition can be either water droplets (low level clouds), ice crystals (high level clouds) or a combination of both phases (mid-level clouds).

Measurements of cloud amount are either made automatically i.e. by satellites and ground-based imagers, or from the ground by visual observation. The total cloud amount, or total cloud cover, is the fraction of the celestial dome covered by all clouds visible. The assessment of the total amount of cloud, therefore, consists in estimating how much of the total apparent area of the sky is covered with clouds (WMO, 2008). According to the World Meteorological Organization (WMO), the proportion of cloud amount is given in eighths or oktas as in Table 2.1 (Castro-Almazán et al., 2015).

Table 2.1: Total cloud amount in oktas and corresponding description of the sky condition. Adapted from Castro-Almazán et al. (2015).

Oktas	Description of sky condition
0	Completely clear
1	Clear
2	Clear
3	Partly cloudy
4	Partly cloudy
5	Partly cloudy
6	Cloudy
7	Cloudy
8	Cloudy

Forecasting of irradiance requires an understanding of its interaction with clouds. As explained by Tapakis and Charalambides (2013), the effect of clouds on solar irradiance is due to factors such as reflection, absorption and scattering of the incoming irradiance by cloud particles, and is strongly dependent on cloud volume, shape, thickness and composition. Furthermore, it is noted that not all clouds have the same effect on irradiance. There exists different cloud types that have different dimensions, opacity and composition properties, and these features result in a different effect on ground-received irradiance. In addition, a single cloud has a different effect on ground-received irradiance as compared to multiple clouds or overcast sky conditions. According to Myers (2013), clouds have a three-dimensional characteristic, making the modeling of solar radiation transfer through them and the atmosphere a very complicated process.

Short-term temporal variability of irradiance caused by clouds and their various characteristics can occur on timescales of seconds to minutes. This type of variability is of particular importance to the solar power station since it causes disturbances in the power output due to shading of all or part of the collector field.

2.4 Clear Sky model

A Clear Sky Model (CSM) estimates ground received irradiance under cloud free skies as a function of various atmospheric parameters, some of which include aerosols, water vapour, ozone, solar altitude angle and altitude of the particular site above mean sea level (Reno et al., 2012). In the absence of ground-based solar monitoring stations, CSMs play a vital role in estimating ground irradiance which is an essential requirement for the deployment of solar power plants. Some clear sky radiation models depend on atmospheric turbidity, commonly known as Linke Turbidity (T_L) (Linke, 1922).

Atmospheric attenuation is caused by the scattering of air molecules and aerosol particles, and by absorption by various atmospheric constituents such as ozone, water vapour, oxygen, and carbon dioxide (Jacovides, 1997). According to Muneer (1997), T_L is defined as the number of clean, dry atmospheres that would produce the same total depletion of direct solar radiation as that produced by the actual atmosphere. It is used to describe the turbidity of the atmosphere, and hence the attenuation of the beam solar radiation fraction and the increase of the diffuse fraction. The larger the T_L, the larger the attenuation of the radiation by the clear sky atmosphere. It depends on the optical thickness of the clean and dry atmosphere which is sensitive to air mass. Therefore, T_L depends on air mass and consequently, on solar elevation (the solar angle describing the height of the Sun above the horizon and compliment of the zenith) at the instant of its evaluation.

Many CSMs require several meteorological and/or atmospheric parameters that are generally inaccessible. Examples of CSMs include Bird, Atwater, MAC and REST2 (Reno et al., 2012). Table 2.2 lists these CSMs and the required atmospheric input parameters.

Table 2.2: Clear sky models and the corresponding atmospheric input parameters required. The required parameters for each model are marked with an "X". Adapted from (Reno et al., 2012).

Parameter	Atwater	Bird	MAC	REST2
Precipitable water	X	X	X	X
Pressure	X		X	X
Ground albedo	X	X	X	
Broadband aerosol optical depth	X	X	X	X
Reduced ozone vertical pathlength	X	X	X	X
Humidity		X		
Temperature		X		
Angstrom's wavelength exponents				X
Aerosol single-scattering albedo	X			X
NO2 pathlength				X
Turbidity			X	X

Atwater and Ball (1981) developed a model that focused mainly on the transmittance of solar radiation through clouds, where the different cloud levels were not distinguished. This differs from the MAC model (Davies and McKay, 1982) which uses cloud information from the different cloud levels. REST2 that was developed by Gueymard (1989) and the model that requires the most number of parameters, is a two-band radiation model that divides the solar spectrum into an ultraviolet/visible and an infra-red band. Within each band, the transmittance of each extinction layer (ozone, water vapor, mixed gases, molecules and aerosols) is parameterized using spectral transmittance functions.

According to Reno et al. (2012), many of the atmospheric parameters presented in Table 2.2 may be estimated using a constant value, however, doing so will decrease the accuracy of the model. One method of acquiring a more accurate representation of these parameters is by employing a full meteorological measurement station at the location of interest to measure all required parameters. However, due to the high cost involved in the setup and operation of these stations, this option may not always be possible. Many of the CSMs require several input parameters that are not easily accessible for Durban. Furthermore, this thesis was restricted to the use of a model that required only ground-measured radiometric data to estimate T_L. Therefore, the Ineichen CSM was chosen.

The advantage of the Ineichen model is that it requires only a single atmospheric parameter i.e. T_L. The T_L values can be estimated using ground-measured beam irradiance data available for Durban, which is recorded using a pyrheliometer as described in Section 2.5.

As described in Ineichen and Perez (2002), Linke (1922) expressed the total optical thickness of a cloudless atmosphere as the product of two terms, δ_{cda}, the optical thickness of a clear and dry atmosphere (water and aerosol free), and T_L which represents the number of clean and dry atmospheres producing the observed extinction according to

$$B_{nc} = I_o \cdot e^{(-\delta_{cda} \cdot T_L \cdot m_A)}, \tag{2.5}$$

where B_{nc} is the beam irradiance, the solar constant I_o = 1367 W/m^2 and m_A is the air mass as a function of the zenith angle θ_z. Given B_{nc}, equation (2.5) provides an estimate T_L. To account for the Earth's curvature, Kasten and Young (1989) approximated m_A as

$$m_A = \frac{1}{cos(\theta_z) + 0.50572(6.07995 + (90 - \theta_z)e^{-1.6364})}. \tag{2.6}$$

Using measured B_{nc}, equation (2.5) can be solved for T_L as

$$T_L = \frac{\ln(I_o/B_{nc})}{\delta_{cda} \cdot m_A}. \tag{2.7}$$

Equation (2.7) enables T_L to be estimated given measured B_{nc} and estimates of I_o, δ_{cda} and m_A. Whereas estimation of I_o and m_A is straightforward, δ_{cda} is more complex. Linke (1922) defined δ_{cda} as the integrated optical thickness of the atmosphere free of clouds, water vapor and aerosols, which was computed from theoretical assumptions and which was validated in a dry mountain atmosphere. For δ_{cda}, Linke (1922) produced the following formulation

$$\delta_{cda} = 0.128 - 0.054 \cdot \log(m_A). \tag{2.8}$$

An alternative formulation by Kasten (1980), was based on a series of spectral data tables published by Feussner and Dubois (1930) which incorporated both molecular scattering and absorption by the stratospheric ozone layer. Kasten (1980) fitted the following equation to these tables

$$\delta_{cda} = (9.4 + 0.9 \cdot m_A)^{-1}, \tag{2.9}$$

which is known as Kasten's pyrheliometric formula. The Kasten Linke turbidity (T_{LK}) is

$$T_{LK} = \ln(I_o/B_n)[9.4 + 0.9 \cdot (m_A)]/m_A, \tag{2.10}$$

where B_n is the measured beam irradiance. Studies by Grenier et al. (1995) and Louche et al. (1986) have proposed the use of updated spectral data in an attempt to improve the formulation of δ_{cda}.

In the determination of a new Linke turbidity coefficient, the approach by Ineichen and Perez (2002) was not to produce a better formulation of δ_{cda}, but rather to use T_{LK} as a reference. Ineichen and Perez (2002) introduced the Linke turbidity coefficient T_{LK} at air mass 2, and a multiplicative coefficient, b, that is dependent on the altitude, a, of the location. For the beam clear sky irradiance, the following empirical expression was obtained

$$B_{ncI} = b \cdot I_o \cdot e^{-0.09 \cdot m_A \cdot (T_{LK}-1)}, \tag{2.11}$$

where $b = 0.664 + \frac{0.163}{f_{h1}}$ and $f_{h1} = e^{\left(\frac{-a}{8000}\right)}$. The value of 8000 refers to the scale height of the Rayleigh atmosphere in meters. The parameter f_{h1} accounts for the altitude of the location with respect to height of the Rayleigh atmosphere. At sea level, i.e. $a = 0$, f_{h1} equals unity and b is approximately 0.83, which is the fraction by which I_o will be reduced. At higher altitudes, $b > 0.83$ and the fractional decrease in I_o is smaller, resulting in a larger fraction of I_o reaching the location. Furthermore, for a clean and dry atmosphere i.e. when $T_{LK} = 1$, there is no atmospheric attenuation and the beam irradiance in equation (2.11) is simply a fraction of the solar constant.

The turbidity according to Ineichen and Perez (2002), and denoted as (T_L), is

$$T_{LI} = \left[11.1 \cdot \ln\left(\frac{bI_o}{B_{ncI}}\right)/AM\right] + 1. \tag{2.12}$$

Equation (2.12) was solved for $m_A = 2$ in order to produce the Kasten (1980) clear sky global irradiance

$$G_{hcK} = 0.84 \cdot I_o \cos(\theta_z) e^{-0.027 \cdot AM(f_{h1} + f_{h2}(T_{LK}-1))}, \tag{2.13}$$

where $f_{h2} = e^{\left(\frac{-a}{1250}\right)}$.

Ineichen and Perez (2002) adjusted the Kasten (1980) model for clear sky global horizontal irradiance to produce

$$G_{hcI} = c_{g1} I_o \cos(\theta_z) e^{-c_{g2} AM(f_{h1} + f_{h2}(TL-1))} e^{0.01 \cdot AM^{1.8}}, \tag{2.14}$$

where $c_{g1} = (5.09 \times 10^{-5})(a + 0.868)$ and $c_{g2} = (3.92 \times 10^{-5})(a + 0.0387)$. The c_{g1} and c_{g2} parameters are the corrections to the altitude of the location, applied by Ineichen and Perez (2002). The coefficients f_{h1} and f_{h2} relate the altitude of the station with the altitude of the atmospheric interactions. At sea level, there will be more atmospheric interaction between the solar irradiance and

atmospheric constituents, as compared to higher altitudes. As suggested by Miller (1981), at higher altitudes there are fewer scattering molecules between the Sun and the Earth's surface.

Similar to equation (2.14), the clear sky diffuse horizontal irradiance (D_{hcI}) may be found by the difference between the global and direct components

$$D_{hcI} = G_{hcI} - B_{ncI} \cdot cos\theta_z. \tag{2.15}$$

For normalization of the irradiance components that are required for clustering, the monthly averages of T_{LI} were estimated using radiometric data in Durban. The radiometric data used in the Ineichen and Perez (2002) model was recorded in Durban during the year 2013 (Zhandire, 2015). Daily records of DNI for Durban used for computing the T_{LI} values can be accessed from the Southern African Universities Radiometric Network (SAURAN) data base (SAURAN, 2014). Examples of daily DNI profiles from this data base are given in Figure 2.4 (a) and (b).

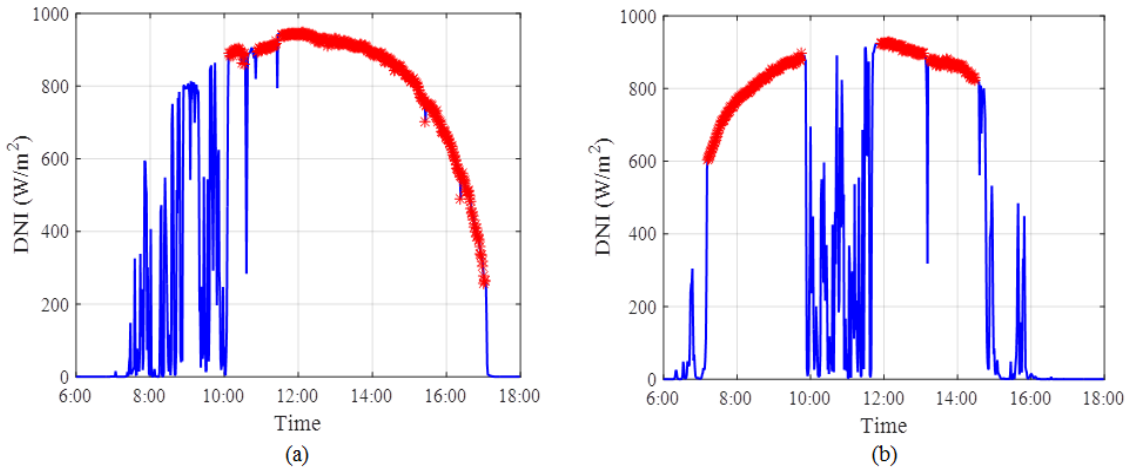


Figure 2.4: An illustration of the selection of clear sky periods from a typical beam irradiance profile during (a) a day with clouds present in the morning and a clear afternoon and (b) a day with clouds present during midday and the late afternoon and clear periods during the morning and early afternoon.

In order to compute the T_{LI} values, only clear sky periods should be used as input into the Ineichen model. The selection clear sky periods for computing the T_{LI} values were chosen according to the criteria described in Ineichen (2006), Chaâbane et al. (2004) and Reno and Hansen (2016). Generally, clear days will have most of the values that satisfy the criteria described in Chaâbane et al. (2004); Ineichen (2006); Reno and Hansen (2016). However, partly cloudy days that have

intermittent periods of clear sky can be extracted and used for estimating T_{LI}. This is illustrated in Figure 2.4 where the clear sky periods (indicated in red) within typical partly cloudy days that occur in Durban, may be extracted and used to estimate T_{LI}. The T_{LI} estimates together with the MATLAB implementation of the Ineichen CSM developed by Sandia National Laboratories (Reno, 2012), produced clear sky estimates of B_{ncI}, D_{hcI} and G_{hcI} referred to as DNI_c, DHI_c and GHI_c, respectively.

Furthermore, to investigate the reliability of the estimated T_{LI} values for Durban, they were compared with turbidity estimates provided Remund et al. (2003), where Linke turbidity maps for the world for each month using a combination of ground and satellite data where produced. This has been made available on the Solar Radiation Data (SoDa) website (SoDa, 2011).The only available turbidity estimates available for Durban from SoDa was for the year 2003. Nevertheless, this was still used to serve as a comparison for the T_{LI} values computed using radiometric data. The month average turbidity estimates for Durban computed for the year 2003, referred to as T_{LS}, are presented in Table 2.3, together with the computed T_{LI} estimates for Durban. For Durban, estimated T_{LI} values range from 2.9 to 3.4 with the lowest and highest values being for July and November, respectively. On the other hand, the estimates for T_{LS} range from 2.9 to 3.9. The annual average T_{LI} and T_{LS} were found to be 3.1 and 3.2, respectively. Overall, the month average T_{LI} estimates computed for Durban can be considered to be fairly reliable since they are relatively consistent with those produced by Remund et al. (2003). According to Reno et al. (2012), the Ineichen and Perez (2002) model was found to have an RMSE of 5%.

The DNI_c, DHI_c and GHI_c profiles are shown in Figures 2.5 and 2.6 for winter and summer solstices, respectively. In Durban, the summer months receive considerably high irradiance levels, where DNI_c almost reaches 1000 W/m^2 and where GHI_c exceeds this value. In winter DNI_c and GHI_c is lower, reaching their maximum of 800 W/m^2 and 600 W/m^2, respectively, at midday.

Table 2.3: Month averages of atmospheric turbidity for Durban for T_{LI} and T_{LS}, with their respective mean. The largest difference in the month average mean T_{LI} and T_{LS} is observed during September. The annual averages for T_{LI} and T_{LS} are relatively consistent.

Month	$\boldsymbol{T_{LI}}$	$\boldsymbol{T_{LS}}$
January	3.2	3.4
February	3.3	3.6
March	3.0	3.3
April	3.1	3.5
May	3.2	3.1
June	3.0	3.0
July	2.9	3.1
August	2.9	2.9
September	3.0	3.9
October	3.4	3.0
November	3.4	3.3
December	3.2	3.0
Average	**3.1**	**3.2**

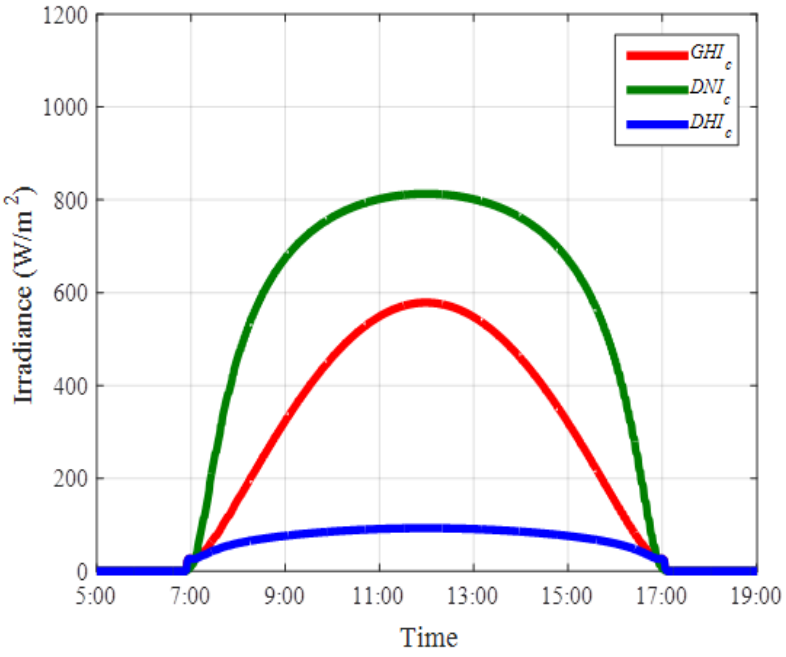


Figure 2.5: Clear sky profiles of direct (DNI_c), diffuse (DHI_c) and global (GHI_c) irradiance for the winter solstice (21 June 2014) in Durban using the Ineichen model (TL = 3.0).

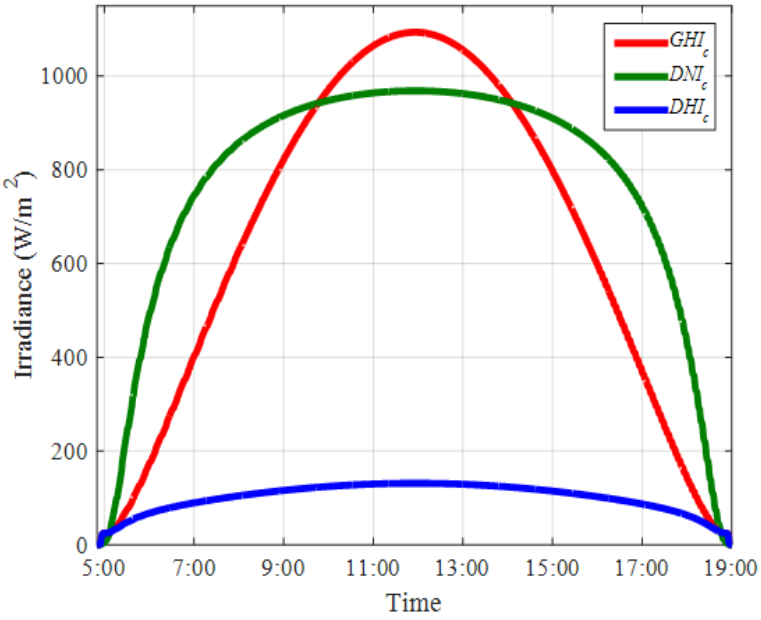


Figure 2.6: Clear sky profiles of direct (DNI_c), diffuse (DHI_c) and global (GHI_c) irradiance for the summer solstice (22 December 2014) in Durban using the Ineichen model (TL = 3.2).

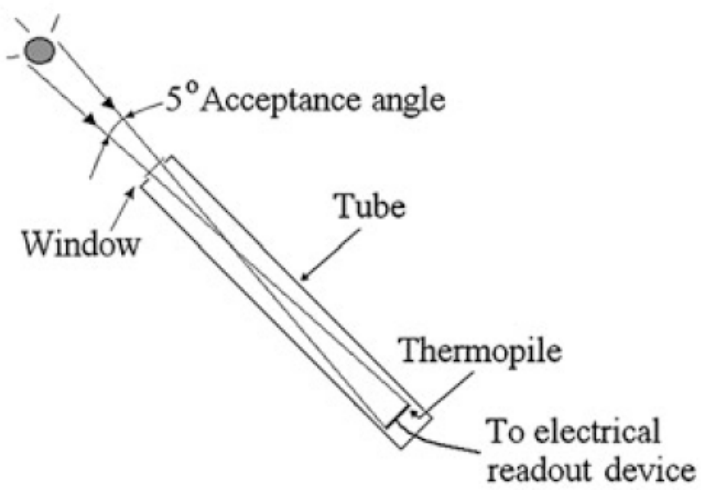
2.5 Solar radiation measurement instruments

The measurement of solar radiation is referred to as *radiometry*, and the purpose of a radiometric detector (or radiometer) is to convert photons of light into a measurable signal (Myers, 2013). Terrestrial solar radiation is measured using instruments such as pyranometers and pyrheliometers. Pyrheliometers measure shortwave direct solar irradiance. It must be pointed at the Sun such that the incoming rays are normal to the optical window and the measured quantity is therefore referred to as direct normal irradiance (DNI) or beam irradiance. Pyranometers measure shortwave global radiation, and diffuse radiation by the use of a shadow band or shading ball across the sensor. Global and diffuse irradiance are referred to as global horizontal irradiance (GHI) and diffuse horizontal irradiance (DHI), respectively. The relationship between global, beam and diffuse irradiance is given in equation 2.1.

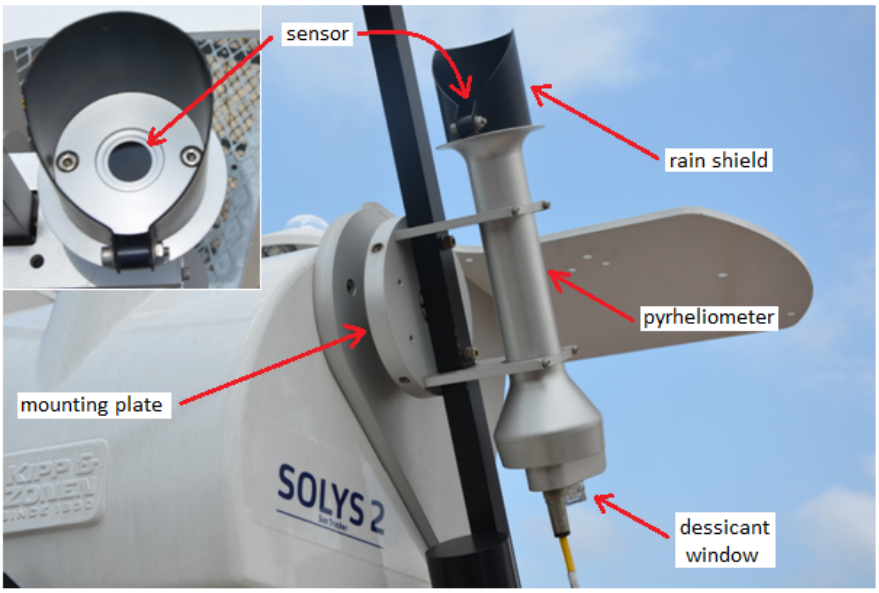
2.5.1 Pyrheliometers

A pyrheliometer is used to measure the direct (beam) solar radiation received at a particular location. Pyrheliometers measure the nearly collimated radiation within a narrow field of view (FOV), typically between 5.0° and 5.8°. The angle of aperture for the instrument used for this study is $5.0^\circ \pm 0.2^\circ$ (Kipp and Zonen, 2014). A schematic of a pyrheliometer is shown in Figure 2.7 (a). The sensing element is placed at the bottom of the tube. When the phyreliometer is pointed at the Sun (detector normal to the direct solar beam), only radiation within the FOV is captured by the detector (Myers, 2005). The sensing element consists of a blackened thermopile that converts heat into an electrical signal. To continuously measure direct solar radiation the pyrheliometer has to be constantly following the Sun and this is achieved by the use of a Sun tracker. For this study, the Kipp and Zonen CHP1 pyrheliometer used to measure the direct normal irradiance was mounted onto the Solys2 tracker, as shown in Figure 2.7 (b).

The radiometery station in Durban is equipped with a Solys2 automatic sun tracker. The tracker has a built-in GPS unit that retrieves information about the location, date and time. The solar angles are then computed using a solar position algorithm. According to WMO stardards, the CHP1 pyrheliometer is of "first class" standard (Kipp and Zonen, 2014). The Absolute Cavity Radiometer (ACR) is of the highest class and all other pyrheliometers are calibrated against it. As reported by Myers and Wilcox (2009), the pyrheliometer has an uncertainty of $\pm 2\%$.



(a)

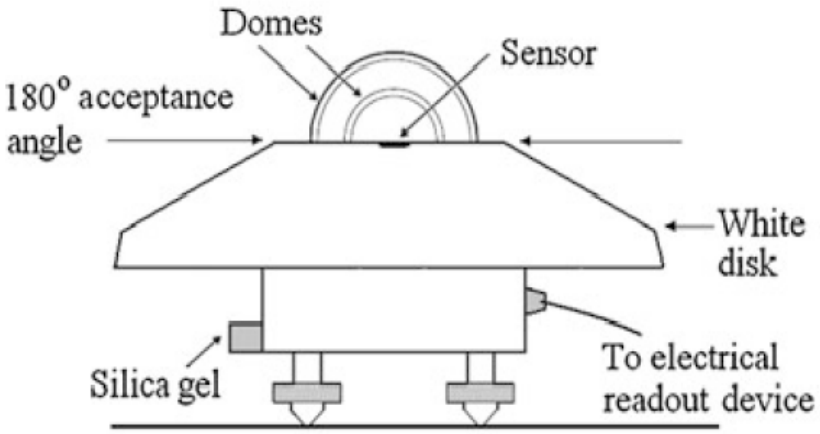


(b)

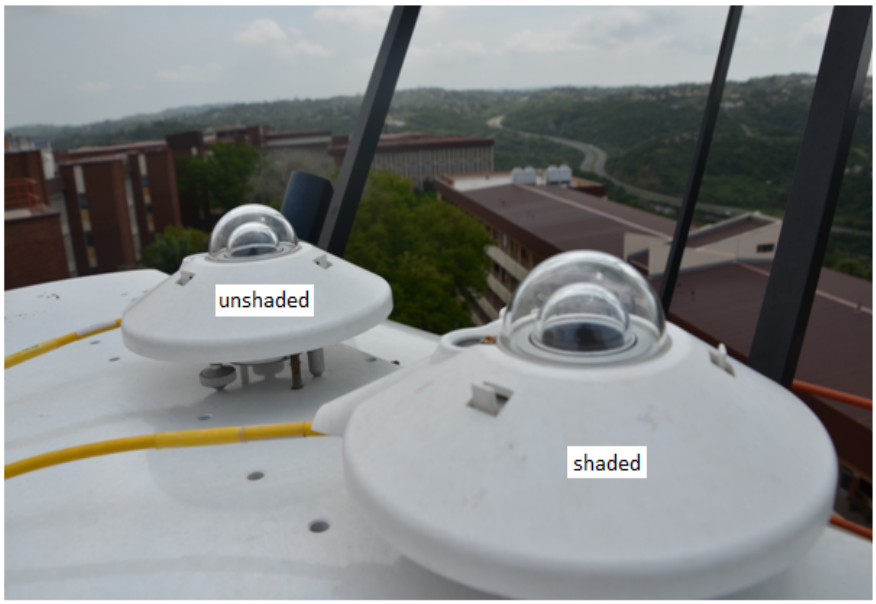
Figure 2.7: (a) Schematic of a pyrheliometer. Only radiation with the narrow FOV is captured by the sensing element situated at the bottom of the tube. Adapted from Paulescu et al. (2013). (b) Kipp and Zonen CHP1 pyrheliometer use to measure the direct normal irradiance, mounted onto the Solys2 tracker.

2.5.2 Pyranometers

Global and diffuse solar irradiance are measured using a pyranometer. Pyranometers measure the radiation that is incident on them within the solid angle 2π. Similar to that of the pyrheliometer, the instrument consists of a blackened thermopile sensing element which is housed in a domed structure. The thermopile converts the temperature to a voltage. The function of the dome cover is to shield the sensing element from wind and rain, as this may affect its temperature. However, the dome still allows transmission of the solar radiation equally from all directions. A schematic of the pyranometer is shown in Figure 2.8 (a). Measurement of the diffuse component can be achieved by the use of a shadow band or shading ball covering the sensing element. The direct component of the solar radiation is blocked by the shading device so that only the scattered and reflected irradiance can be received by the sensor. Alternatively, if the direct component is known the diffuse component can be calculated using equation (2.1). For this thesis, Kipp and Zonen CMP11 shaded and unshaded pyranometers used to measure the diffuse and global components respectively, and are shown in Figure 2.8 (b). As reported by Myers and Wilcox (2009), the pyranometer has an uncertainty of $\pm 5\%$.



(a)



(b)

Figure 2.8: (a) Schematic of a pyranometer. Adapted from Paulescu et al. (2013). (b) Shaded and unshaded pyranometers mounted on a stationary plate on top of the Solys2 tracker, at the Howard College radiometry station in Durban. The diffuse component is measured by the shaded pyranometer, which uses a shadow ball to cover the sensor.

The radiometry station consisting of the above-mentioned instrumentation is located in Durban at the Howard College campus of the University of KwaZulu-Natal (S 29.871 °; E 30.977 °) 150 m above mean sea level, on a roof platform that has a largely unobstructed view of the horizon. Figure 2.9 shows the radiometry station. The instruments are connected to an 8-channel Campbell Scientific CR1000 data logger, and measurements of DNI, DHI and GHI are sampled at 0.5 Hz, and averaged for minute, hourly and daily intervals. A solar panel at the site provides power to the instrumentation and ensures continuous data collection. Instruments are also subject to regular maintenance to ensure high quality measurements.

The Howard College station is part of the Southern African Universities Radiometric Network (SAURAN) which is a national radiometric network that was established in 2014 in South Africa. The network consists of 22 stations and aims to make high-resolution, ground-based solar radiometric data available from stations located across the Southern African region, including South Africa, Namibia, Botswana and Reunion Island (Brooks et al., 2015).

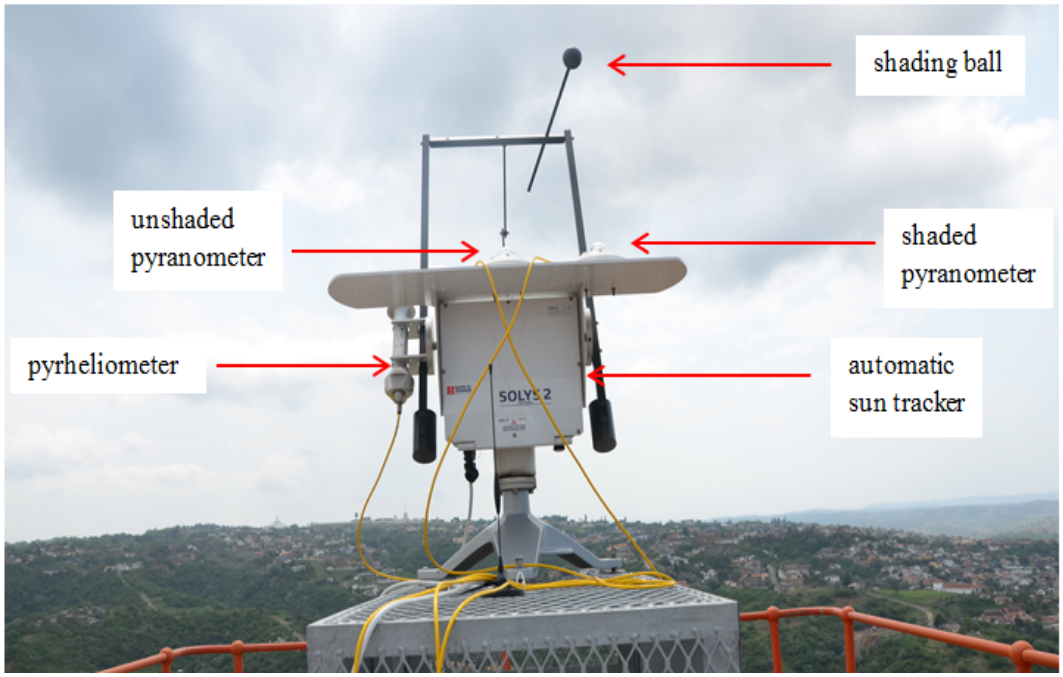


Figure 2.9: Instrumentation at the radiometry station in Durban. The pyrheliometer, shaded and unshaded pyranometers are mounted onto a Solys2 sun tracker.

2.6 Data and definition of variables

The radiometric data used for this study were daily irradiance profiles in Durban that were recorded from 28 January 2014 to 27 January 2015 during 8:30-16:30 solar time. This set of profiles was used for the clustering analysis. Although this interval is not symmetric about solar noon, it does not affect the clustering and forecasting methods. Testing of the forecasting methods were done using a set of 100 days in 2017, including most days during the months of January to June. As will be described in Chapter 6, the test set of 100 days were recorded at the same time as the cloud cover forecasts from the NWP. Minute-resolution profiles of the DNI, DHI and GHI components were recorded and from here on, they are referred to as *B*, *D* and *G*, respectively. Here, the clear sky equivalents are denoted as B_c, D_c and G_c.

As discussed in Chapter 4, clustering of minute-resolution horizontal beam irradiance fraction, k_b, was used as an example to illustrate the pre-processing and clustering methods. Although the present work is similar to that of Jeanty et al. (2013) where daily k_b profiles were clustered, the present work instead applies clustering to the normalized beam irradiance for reasons outlined in Chapter 5. Nevertheless, for completeness of definition of variables, k_b is defined here as

$$k_b = 1 - \frac{D}{G}, \tag{2.16}$$

where *G* must be greater than zero. The value of k_b ranges from 0 to 1. k_b close to 0 indicates cloudy sky conditions. Alternatively, a k_b value close to 1 indicates sunny sky conditions.

In this study, the normalized irradiance components will be used for clustering. To obtain the minute-resolution normalized beam irradiance, B_n, the beam irradiance profiles, B, were normalized to the beam component, B_c of the Ineichen clear sky model as follows:

$$B_n = \frac{B}{B_c}. \tag{2.17}$$

In a similar manner, the minute-resolution normalized diffuse irradiance D_n is

$$D_n = \frac{D}{D_c}. \tag{2.18}$$

The hour average of B_n, denoted as $\bar{B}_n$, is

$$\bar{B}_n = \frac{1}{p}\sum_{i=1}^{p} B_{ni}, \tag{2.19}$$

where *i* is the minute index and *p* = 60. The first minute (*i* = 1) is on the half-hour, e.g. 8:30 so values of $\bar{B}_n$ are known at times 9:00, 10:00, 11:00,...16:00. Similarly, the hour average of D_n, denoted as $\bar{D}_n$, is

$$\bar{D}_n = \frac{1}{p}\sum_{i=1}^{p} D_{ni}. \tag{2.20}$$

In addition to clustering irradiance profiles, profiles of the variability in the irradiance were also clustered. The variability in the minute-resolution B_n is given by

$$V_B = \sqrt{\frac{1}{p}\Big(\sum_{i=1}^{p} B_{ni} - \bar{B}_n\Big)^2}, \tag{2.21}$$

where *p* = 60 and therefore V_B is the variability over the hour.

For this thesis, normalization does not specifically refer to the mapping of the data onto a [0, 1] range. Instead, D_n was normalized to the clear sky value so that it could be used as a reference, to allow for the comparison between D_n for different days due to the presence of clouds. Therefore, for a clear day D_n will be 1 and for cloudy days it will exceed 1. Figures 2.10 and 2.11 show examples of B_n and corresponding D_n profiles for a highly variable day. B_n can vary considerably within the hour and D_n exceeds 1, depending on the amount of cloud cover.

The single-point measurements using the radiometer serve as an indicator of conditions over the whole sky. The relationship between irradiance and amount of cloud cover is described by Lam and Li (1998) as, “less cloud cover means a clearer sky, and hence more solar radiation”. Therefore, for the present work, it is assumed that the movement of clouds results in an irradiance signal that varies with time and is correlated with cloud cover conditions.

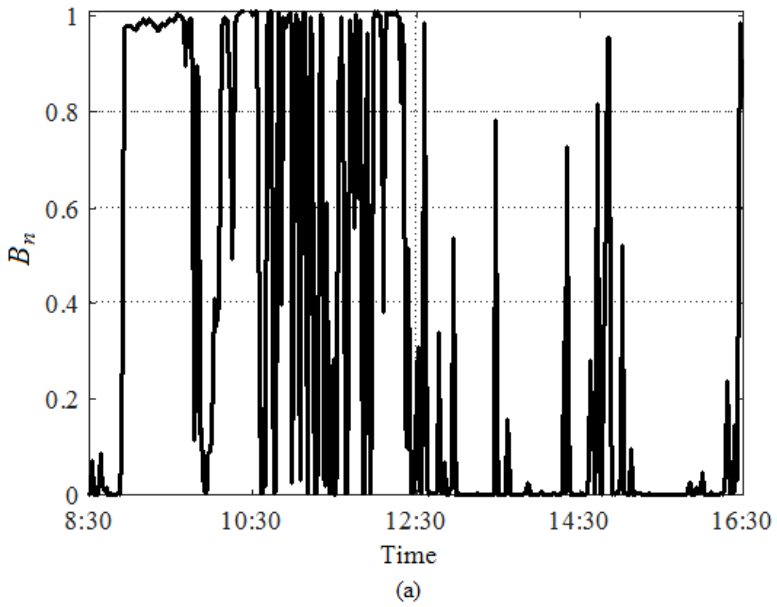


Figure 2.10: B_n profile for a highly variable day (17 January 2017) in Durban where B_n varies considerably depending on the amount of cloud cover.

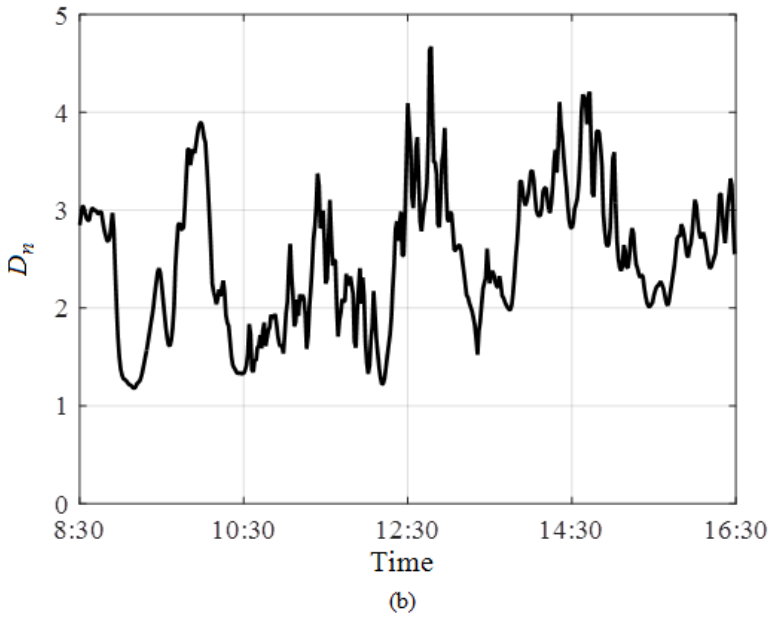


Figure 2.11: D_n profile for a highly variable day (17 January 2017) in Durban where D_n exceeds 1 for cloudy conditions.

Chapter 3

Overview of solar forecasting methods

Different forecasting methods for different forecast horizons, from intra-day to several days ahead, have been proposed (Marquez and Coimbra, 2013). These include statistical (linear and non-linear), image-based, which involve the use of satellites and ground-imagers, and Numerical Weather Prediction (NWP). An overview of these methods, their performance and the forecast horizon within which they are most effective, are discussed.

3.1 Forecasting methods

Solar forecasting employs different methods that can be broadly categorized as statistical-based, image-based and NWP-based. These methods make use of either time series irradiance measurements, images of cloud cover from satellites or ground-based sensors, or meteorological variables as inputs into mathematical models. The success of the forecasting method also depends on the forecast horizon, which is the amount of time in the future for which the forecast is prepared. Depending on the forecast horizon, some methods perform better than others.

The objective of this thesis was day-ahead forecasting using clustering of irradiance profiles. The approach taken was the combination of cloud cover forecasts from an NWP with irradiance classes produced by clustering, for day-ahead forecasts of normalized beam and diffuse irradiance. As will be discussed, other methods of forecasting such as application of CMVs to cloud imagery from satellites and ground-based imagers, ARIMA and ANNs were not used. These methods either require high temporal and spatial resolution imagery that is expensive to acquire, have a limited forecast horizon or require large amounts of data for training and model development.

Figure 3.1 gives a summary of the relationship between the different forecasting methods and their corresponding spatial and temporal resolution within which they perform best. In general, statistical methods are best for short time horizons, NWP methods for long time horizons and satellite methods for the intermediate. Each of these methods is discussed as well as the combination of two or more methods, known as hybrid forecasting methods.

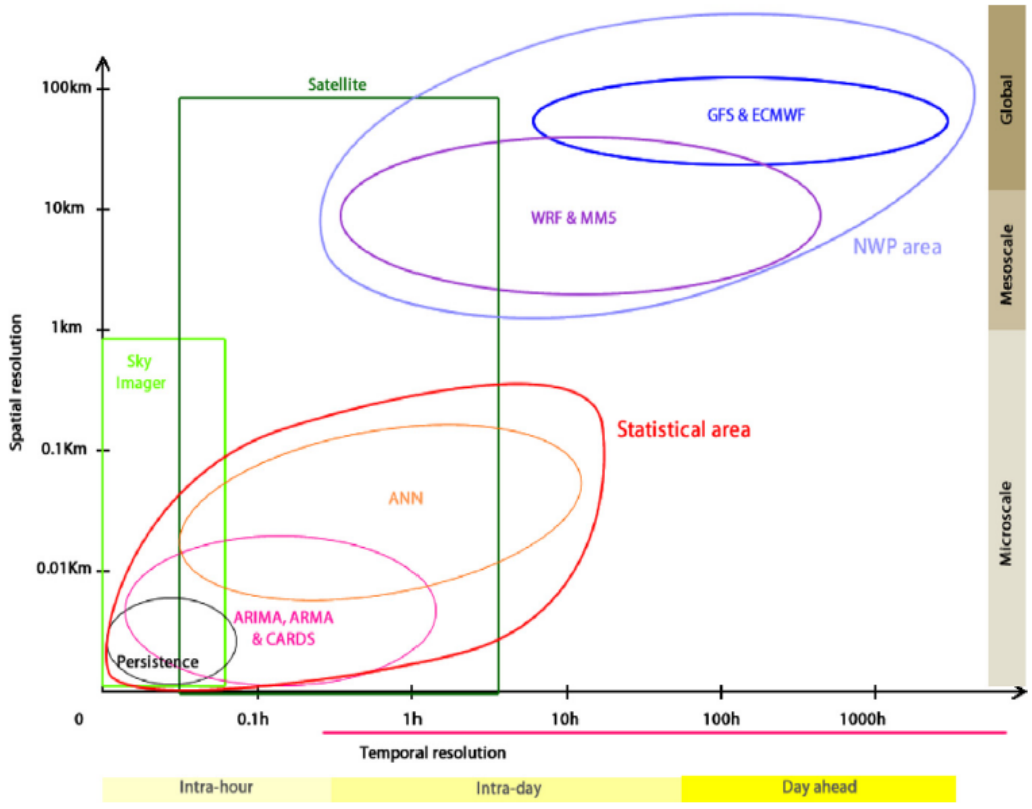


Figure 3.1: Summary of the relationship between the different forecasting methods and their corresponding spatial and temporal resolution. Persistence and sky imagery are only applicable in the intra-hour range. Statistical models such as Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA), Artificial Neural Network (ANN), as well as satellite imagery span the intra-hour and intra-day ranges. Day-ahead forecast range is dominated by global and mesoscale NWP models. Adapted from Diagne et al. (2013).

3.1.1 Statistical

Statistical forecasting methods involve the use of historical data and can be broadly categorized as linear models or non-linear models. Techniques in both these categories aim to establish relationships between predictor variables and the variable to be predicted. Examples of linear statistical models are ARMA, ARIMA, Multiple Linear Regression (MLR) and Exponential Smoothing (ES). Alternatively, non-linear models include genetic algorithms (GAs), ANNs and wavelet neural networks (WNNs) (Diagne et al., 2013). Voyant et al. (2017) presents a review of several statistical and machine learning techniques that are applied to solar radiation forecasting.

3.1.2 Image-based

Solar forecasts using imagery can either be from satellites or ground-based sky cameras. An image from the visible (VIS) band of the satellite is a photograph of a specific region of the Earth that captures the reflected sunlight from the upper surface of the clouds (Tapakis and Charalambides, 2013). Assuming cloud features remain constant and applying CMVs, future cloud positions (based on their trajectory and velocity) are predicted and the amount of radiation reaching the ground can be estimated. An example of a satellite cloud image with the CMVs applied is shown in Figure 3.2. This method allows for accurate forecasts up to 6 hours ahead (Kostylev and Pavlovski, 2012). The spatial scale covered by satellite images is much larger as compared to ground-based images. As a result, satellite images are better suited to providing forecasts for time horizons of more than 3 hours ahead.

Sky images acquired using a ground-based sky camera, such as in Figure 3.3, offer a more detailed picture of the cloud extent, structure and motion. However, their field of view is reduced and their forecast horizon is limited to less than an hour (Pelland et al., 2013). Nevertheless, their high temporal resolution (minute and sub-minute) makes them useful in the prediction of ramp rates, defined by Kleissl (2013) as the change in irradiance over some time, as individual cloud motion tracking is possible (Mathiesen, 2013).

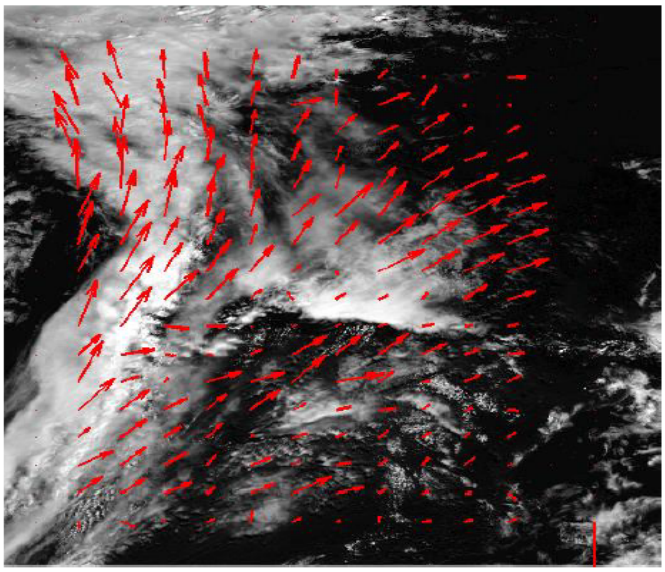


Figure 3.2: Cloud motion vectors applied to a satellite image. Adapted from Cros et al. (2014).

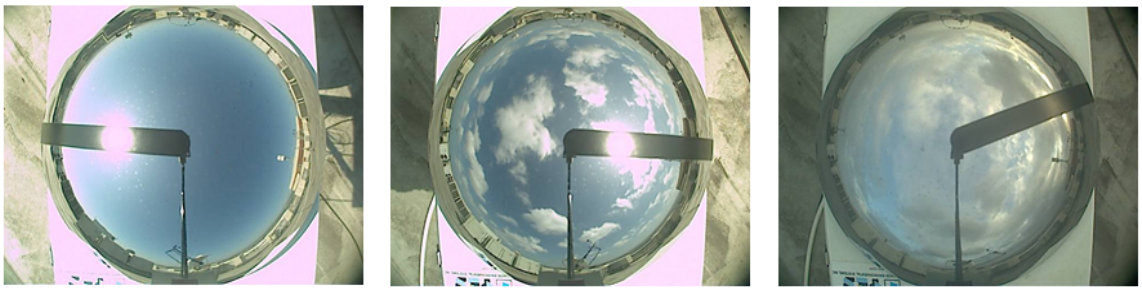


Figure 3.3: Sky images taken with a total sky imager (model TSI-440) in Durban on 24 December 2015. (a) clear conditions at 8:44 (b) partly cloudy conditions at 14:21 and (c) cloudy conditions at 18:00.

3.1.3 Numerical Weather Prediction

Numerical Weather Prediction (NWP) relies on mathematical models to predict the evolution of the atmosphere based on initial conditions. NWP models are generally categorized by their domain i.e. global or regional. The initial atmospheric conditions required as input are obtained from satellite, radar, radiosonde and a large network of ground measuring stations (Pelland et al., 2013). NWP models provide a variety of outputs, including temperature, humidity, precipitation, wind speed and direction, and cloud cover. It is cloud cover that will be used in the present study because of its

strong association with beam irradiance intensity.

One of the most commonly-used global NWP models is the Global Forecast System (GFS), which is operated by the National Oceanic and Atmospheric Administration (NOAA). The GFS model has a spatial domain of 28 km x 28 km, is run every 6 hours to produce forecasts up to 180 hours (7.5 days) ahead and every 12 hours for forecasts up to 384 hours (16 days) ahead. In addition to the 28 km x 28 km horizontal discretization, the GFS models 64 vertical layers of the atmosphere. The coarse spatial resolution of NWP models limit them to large-scale atmospheric conditions. As a result, NWP models are unable to resolve micro-scale conditions associated with cloud formation (Chaturvedi, 2016; Inman et al., 2013). For solar forecasts, this is considered to be one of the largest sources of error when using NWP models. In an attempt to overcome the issue of coarse spatial resolution, regional models are used as an alternative to global models. Regional models such as North American Mesoscale (NAM), Weather Research and Forecasting (WRF), Rapid Update Cycle (RUC) and High Resolution Rapid Refresh (HRRR), cover a limited spatial domain with greater detail and provide a better characterization of the cloud cover conditions which are required for solar forecasts.

For this thesis, cloud cover forecasts for Durban from an NWP output were combined with classes produced by clustering to forecast the normalized beam and diffuse irradiance profiles, for the day ahead. Cloud cover forecasts were obtained from AccuWeather, a public weather- service provider that uses the GFS model to produce hourly-resolution forecasts of cloud cover for the day ahead.

3.1.4 Hybrid methods

Hybrid models are a combination of two or more of the previously-described forecasting methods. These can also be referred to as combined models or ensemble models (Diagne et al., 2013). A motivation for the development of a hybrid model is that often it is possible to increase the forecasting accuracy by taking advantage of the strengths of each methodology (Inman et al., 2013). It is clear that no individual forecasting methodology is able to span all necessary spatial and temporal resolutions. Therefore, by combining two or more methods, forecasts on several spatial and temporal resolutions may be achieved.

With the intention of increasing forecast accuracy, Cao and Cao (2005) and Cao and Cao (2006) developed a hybrid model, which was a combination of an ANN with wavelet analysis, to forecast

total daily irradiance. Similarly, Cao and Lin (2008) used meteorological observations as inputs to a model that combines an ANN with wavelets, to predict hourly GHI values.

Voyant et al. (2012) proposed a model that combined ARMA and ANN (in particular multi-layer perceptron (MLP)) to forecast the hourly GHI for five places in the Mediterranean area. In another study, Marquez and Coimbra (2013) developed a model that combined stochastic learning methods, ground experiments and the National Weather Service (NWS) database, to forecast global and direct irradiance. More recently, Aguiar et al. (2016) combined ground measurements with NWP and satellite data inputs to improve intra-day solar forecasting of GHI.

3.2 Forecast horizon

Before discussing the above-mentioned forecasting methods and the temporal resolutions within which they perform the best, three forecast horizons are defined. The forecast horizons defined by Kostylev and Pavlovski (2012) and Diagne et al. (2013) are as follows:

- Intra-hour: 15 minutes to 2 hours ahead with 30 second intervals to 5 minute intervals.
- Intra-day: 1-6 hours ahead with hourly intervals.
- Day-ahead: 1-3 days ahead with hourly intervals.

Figure 3.4 illustrates the relationship between the forecasting horizons, the forecasting method or models most appropriate for each horizon and the relevant industry related activity.

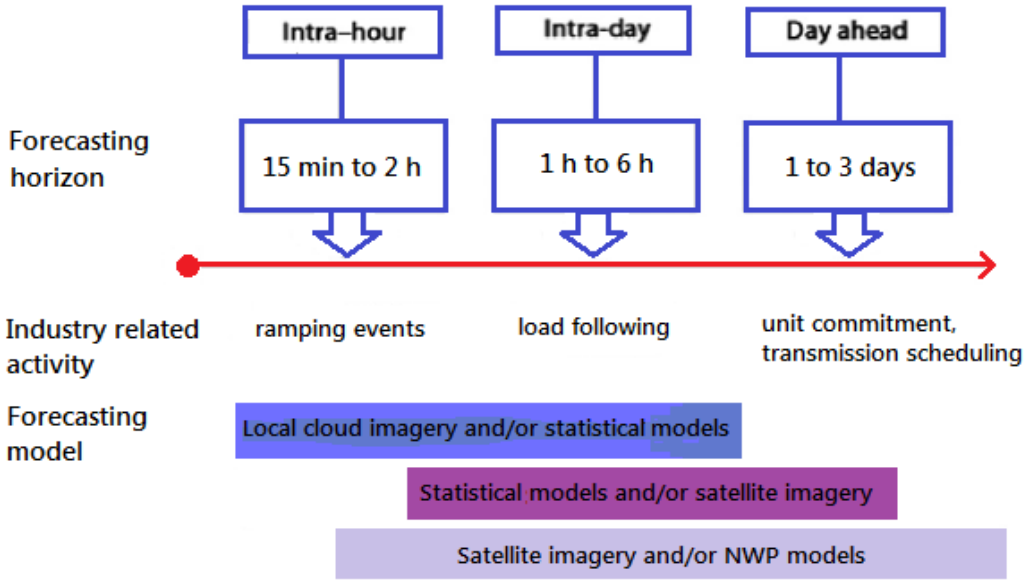


Figure 3.4: Relationship between forecasting horizon, methods and the activity related to industry. Adapted from Diagne et al. (2013).

3.2.1 Intra-hour forecasts

Forecasts in this range are defined by a temporal resolution of 15 minutes up to 2 hours and, in industry, are mainly related to ramping events (Kamath, 2010). The simplest forecasting technique, Persistence Forecasting, also known as the Naive Predictor, is based on the assumption that the forecast, x, at time t+1 is the same as at t,

$$x_{t+1} = x_t. \tag{3.1}$$

It is worthwhile to implement a complex forecasting technique if it is able to improve on Persistence. The accuracy of a Persistence Forecast decreases significantly with forecast duration and, in general, is not an accurate forecast technique for time horizons exceeding 1 hour. Therefore, Persistence Forecasts should only be used as a benchmark or a baseline to compare with more advanced techniques (Diagne et al., 2013).

In order to evaluate forecast performance the three most commonly used metrics are Root Mean Square Error (RMSE), Mean Bias Error (MBE) and Mean Absolute Error (MAE). The RMSE is a

measure of the average spread of the errors and is given by the following equation

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{measured,i} - x_{predicted,i})^2}, \tag{3.2}$$

where $x_{predicted,i}$ and $x_{measured,i}$ represent the *i*th valid forecast and observation pair, respectively and *n* is the number of evaluated data pairs. The MBE is a measure of the average bias of the model expressed as

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(x_{measured,i} - x_{predicted,i}), \tag{3.3}$$

where MBE can be either negative (forecast is too small, on average), zero (forecast has no bias), or positive (forecast is too large, on average) (Diagne et al., 2013). The MAE given by

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \mid x_{measured,i} - x_{predicted,i} \mid, \tag{3.4}$$

and considers only the absolute value of the errors. Of these metrics, RMSE is most commonly reported when describing forecast accuracy (Inman et al., 2013).

A method of forecasting GHI up to 1 hour from HelioClim-3 images was proposed by Dambreville et al. (2014). To estimate future GHI values, CMVs were obtained using a block matching algorithm applied to successive cloud images. It was found that the relative RMSE for the 15, 30 and 45 minute time horizons remained below 30%, which was lower than the Persistence Forecast. This indicated that the block matching algorithm approach was a significant improvement over Persistence Forecasts, especially at longer time horizons.

Similar to the study by Dambreville et al. (2014), Hammer et al. (1999) also used satellite images in the forecast of solar radiation, but a statistical approach to derive the CMVs was employed. Predictions were made for timescales from 30 minutes up to 2 hours. For the 30 minute and 2 hour forecast horizon, RMSE values of 17 and 30%, respectively, were achieved. In addition, a reduction between 7%-10% in RMSE for Persistence Forecasts was found.

In an attempt to apply a more advanced image processing technique, Cros et al. (2014) proposed a forecasting method based on a 2D Fourier transform phase correlation algorithm for cloud motion estimation. This method was found to be significantly improved over the Persistence Forecast, but showed only slight improvement over the traditional block matching algorithm. However, it was found that the computation time involved in the phase correlation method was approximately 25 times faster than previous methods.

The main advantage of satellite technology is that a large area of the Earth can be imaged at once due to the large field-of-view (FOV), therefore providing large-scale cloud information. However, these images are of fairly coarse spatial resolution and of low temporal resolution. In addition, satellites are unable to detect multi-layered clouds since they are only able to image the highest cloud deck and clouds below this will be undetected. Due to this, low clouds contributing to local weather conditions are not detected. Lastly, there is a high cost involved in the acquisition of measurements from the satellite operator (Tapakis and Charalambides, 2013).

As an alternative, ground-based imagers can be employed to overcome some of the problems associated with satellite imaging. Ground-based imagers are able to provide cloud information at high spatial and temporal resolutions (which is not possible with satellites), and more recently, have been used to produce intra-hour forecasts. The cross-correlation of consecutive sky images from a ground-based imager provides a method of monitoring the direction of cloud movement and estimation of their velocities. Irradiance is predicted for the current cloud shadow and then the cloud shadow is moved forward in time based on cloud velocity and direction. The method assumes persistence in the opacity, direction, and velocity of movement of the clouds (Pelland et al., 2013).

A ground-based imager typically consists of a charge-coupled device (CCD) camera that points upwards and photographs the state of the sky at scheduled intervals (Tapakis and Charalambides, 2013). Ground-based imagers are able to provide local cloud information such as cloud cover and cloud fraction, but they are not able to provide information on cloud base height (CBH) and cloud type. An example of a commercially-available ground-based imager is the Total Sky Imager (TSI) manufactured by Yankee Environmental Systems (YES). Ground-based imagers were initially developed for the purpose of monitoring local sky conditions (Calbó et al., 2001; Calbó and Sabburg, 2008; Cazorla et al., 2008; Huo and Lu, 2009; Johnson et al., 1989; Kassianov et al., 2005; Kazantzidis et al., 2012; Long et al., 2006; Martínez-Chico et al., 2011; Pfister et al., 2003; Shields et al., 1993, 2013) on a continuous basis and in the case of cloud assessment and characterization, to provide an alternative to the high cost associated with human observers.

More recently, however, their application has been extended to solar forecasting. For example, Chow et al. (2011) developed a method of forecasting cloud movement and irradiance using a ground-based sky imager. Cloud motion vectors were generated by the cross-correlation of two consecutive sky imagers. Depending on the cloud height and speed, the authors suggest this method is suitable for predicting GHI for the forecast horizon in the range of 5-25 minutes. Some of their

future work includes employing additional sky imagers to increase coverage area and forecast horizon, collocation of a ceilometer (an instrument for measuring cloud base height) with the imager for cloud height retrieval and combining the present method with satellite and NWP-based forecasts to yield a more comprehensive forecast method.

Rather than forecasting GHI as in the work of Chow et al. (2011), Marquez and Coimbra (2013) used TSI imagery to produce intra-hour forecasts of direct normal irradiation (DNI). Their method employed a technique known as Particle Image Velocimetry (PIV). As described by Abdulmouti and Mansour (2006), PIV is a technique used for analyzing two or three-dimensional complex flow fields, and is based on fluid visualization and image processing techniques. The flow of a fluid can be visualized by seeded particles. The distances moved by particles in a flow field can be measured from a series of consecutive images to calculate their speed and direction of the fluid flow. PIV produces a velocity vector map that describes the velocity vector field, and this could be used to extract further physical information such as the pressure field and the vorticity field. When applied to cloud images, the PIV technique produces a map of the cloud velocity field such that the future cloud positions may be estimated in order to forecast irradiance. Marquez and Coimbra (2013) applied the PIV technique to a sequence of TSI images taken at 1 minute intervals to obtain cloud velocity fields. This together with estimating grid cloud fractions in an area of interest was used to forecast DNI. When compared to the persistence model it was found that the optimal forecast period was 5 minutes ahead. In accordance with the work of Chow et al. (2011), the authors demonstrated that the TSI is useful for predictions up to 15 minutes ahead, with the lowest error associated with the 5-6 minute time horizon.

A novel approach used for cloud tracking applied to TSI images was introduced by Quesada-Ruiz et al. (2014). Termed the "sector-ladder" method, the process involves first identifying cloud motion direction and thereafter a size-adjustable set of grid elements was used to assess DNI for 1 minute up to 20 minutes ahead. This method showed a reduction in DNI RMSE under different sky conditions i.e. broken-sky, clear-sky and overcast days. Compared to the PIV method used by Marquez and Coimbra (2013), the sector-ladder method showed an improvement in computational time.

The disadvantages associated with ground-based imagers relate to systematic detection errors such as misdetection of thin clouds and limitations in distinguishing cloud types (Tapakis and Charalambides, 2013). In addition, systems such as the TSI have limited FOV, thus limiting cloud

information to a small area. Due to this, the maximum forecast horizon using a TSI is restricted to 30 minutes. A potential solution to extending the time horizon is to distribute an array of imagers in order to obtain more information on the local cloud situation. Additionally, to obtain information regarding the cloud height, a ceilometer can be co-located with the imager. Despite their shortcomings, these instruments provide on-site measurements of local cloud variation at high spatial and temporal resolution, and due to this unique feature they will continue be a useful tool for solar forecasting (Inman et al., 2013).

Reikard (2009) applied an ARIMA model to six GHI data sets at 5, 15, 30 and 60 minute resolutions. A comparison was made with other methods such as transfer functions, hybrid models and the ANN method. It was found that in most cases, the best results were obtained using the ARIMA model. In addition, the author points out that the ARIMA model has the ability to capture transitions in irradiance associated with the diurnal cycle more effectively than the other methods.

The present work focuses on forecasting for the day-ahead. Given the lack of availability of ground-based sky imagery in Durban, their use was not pursued for this thesis. Furthermore, their limited forecast horizon (i.e. sub-hourly) would not be effective, for this investigation, since clustering produces classes that are based on diurnal irradiance patterns.

3.2.2 Intra-day forecasts

The forecasts in this category that are in the range of 1-6 hours are relevant to load-following and variability associated with operations for grid connected solar powered systems. Some of the previously-discussed methods are also applicable to the intra-day forecasting range. For example, forecasting based on CMVs obtained from satellite images performs well in the temporal range of 30 minutes to 6 hours (Diagne et al., 2013), thus incorporating both, the intra-hour and the intra-day forecast ranges.

NWP models are also used as a forecasting tool for the intra-day forecasting range. For example, Remund et al. (2008) evaluated three NWP-based GHI forecasts (ECMWF, NDFD and GFS/WRF) in the USA. The NWP MBEs were compared and ECMWF showed the best results, followed by NDFD and GFS/WRF. The GFS next day GHI forecasts had an MBE of 19%. This MBE was found to be approximately constant for intra-day (hour-ahead) forecast horizons.

Using METEOSAT data, Lorenz et al. (2004) applied motion vector fields to forecast cloud index images. Comparison with ground data showed a significant improvement in forecast accuracy

when compared to satellite and ground-derived Persistence.

Studies incorporating non-linear statistical methods include Crispim et al. (2008) and Ferreira et al. (2012). The first study made use of artificial intelligence techniques and cloudiness indices obtained from pixel classification of TSI images. The cloud indices were used as inputs to an ANN, optimized with a GA, for prediction of solar irradiance. Similarly, Ferreira et al. (2012) developed their own portable sky imager and created ANN models optimized via a multi-objective GA to predict GHI, cloudiness and temperature for forecasting horizons up to 4 hours.

For intra-day forecasts with a time horizon more than an hour, ground-imaging systems such as the TSI find limited application due to the restricted time of approximately 30 minutes for which clouds are within the FOV (Inman et al., 2013). Nevertheless, the aforementioned methods are successful in providing good forecasts within the intra-day range.

3.2.3 Day-ahead forecasts

For industry-related purposes, day-ahead forecasts are required for operational planning, programming backup, maintenance and transmission scheduling, as well as for planning of reserve usage and peak load matching. From 6 hours up to several days ahead solar irradiance forecasts rely mainly on NWP models.

For the present work, day-ahead forecasts were the focus and particularly how the combination of clustering of irradiance from ground-based instruments and cloud cover forecasts from the NWP could be used to produce a day-ahead forecast of irradiance. The classes produced by clustering describe the diurnal irradiance patterns that can be correlated with diurnal cloud cover patterns. Cloud cover forecasts from the NWP is easily accessible through AccuWeather, thus meeting the need for high spatial and temporal resolution satellite images that are expensive and difficult to obtain. Furthermore, AccuWeather provides a day-ahead forecast that is appropriate for combination with clustering, since clustering of irradiance profiles finds patterns that is on the diurnal-scale. Considering this, the present study focused on day-ahead forecasts of irradiance for Durban.

An evaluation of three NWP forecasts i.e the North American Model (NAM), Global Forecast System (GFS) and European Centre for Medium-Range Weather Forecasts (ECMWF) for predicting global irradiance was conducted by Mathiesen and Kleissl (2011). For all models, MBE and RMSE exceeded 30 and 110 W/m^2, respectively. A general under-prediction in cloud cover was also observed.

Using the National Digital Forecast Database (NDFD), Perez et al. (2007) derived surface global irradiance from the sky cover product to produce a forecasting model. Results showed that for 8-26 and 26-76 hour forecast horizons, the relative RMSE for hourly-averaged global irradiance was 38 and 40%, respectively.

Remund et al. (2008) evaluated three NWP-based GHI forecasts (ECMWF, NDFD and GFS/WRF) in the USA, reporting relative RMSE values ranging from 20% to 40% for the day ahead forecast horizon. Similar results were reported by Perez et al. (2010), where NWP-based irradiance forecasts in several places in the USA were evaluated.

Currently, NWP models are unable to predict the position and extent of cloud fields precisely. This is primarily due to their relatively coarse spatial resolution (1-20 km), rendering them inefficient at resolving micro-scale physics associated with cloud formation. However, NWP models have the advantage of producing forecasts over long time horizons (15-240 hours) and have been shown to be more accurate than satellite-based models for time horizons exceeding 4 hours (Inman et al., 2013). Mesoscale NWP models (such as MM5 and WRF) have higher spatial and temporal resolution and may provide better accuracy in resolving cloud. However, even mesoscale models may not be able to capture cloud movement on the short time scales required at solar power plants, since their output is generally hourly.

Prediction of solar irradiance for more than one day ahead is not restricted to NWP-based models, but can also include statistical techniques. An example of such a study that was undertaken by Martin et al. (2010). The authors presented a comparison of linear and non-linear statistical models applied to half daily values of global solar irradiance with a temporal horizon of 3 days. It was found that the neural network model yielded the best results. Other studies using similar techniques include Paoli et al. (2010, 2014). Techniques using ANNs and other artificial intelligence methods for modelling and forecasting of solar irradiance are presented by Mellit (2008).

It is evident from this discussion that certain forecasting techniques are successful within certain time horizons. Therefore, the choice of forecasting model is strongly dependent on two factors: (i) the forecast horizon and (ii) the available data at a particular site. Ground-based imagers have a maximum forecast horizon of approximately 30 minutes. Statistical methods have been successfully applied to forecast solar irradiance for time horizons ranging from several minutes to a few hours ahead. Satellite imaging is most accurate in producing forecasts up to 6 hours ahead. Forecasts beyond the 6 hour time horizon and up to several days ahead are most accurate if derived from

NWP models.

The focus of this research was the use of clustering for classification and forecasting of irradiance for the day-ahead. Clustering was first used to understand and classify the solar irradiance patterns in Durban. These classes have mean profiles that describe the diurnal irradiance patterns. The classes produced by the clustering were then combined with cloud cover forecasts from the NWP to forecast an irradiance class for the day ahead.

Chapter 4

Cluster analysis

Cluster analysis is a technique used for exploring and identifying interesting patterns and distributions, and discovering natural groupings within data. This chapter reviews previous studies on clustering, and thereafter focuses on the details of two clustering techniques i.e. hierarchical clustering and *k*-means clustering and explains the purpose of each. Pre-processing of the data that is applied prior to the clustering techniques is also discussed. The minute-resolution horizontal beam irradiance fraction is used an example to illustrate all of the above-mentioned techniques.

4.1 Clustering of irradiance patterns

Accurate time series solar radiation data at a given location are vital for the design and deployment of solar energy systems. In addition, time series data are also used for monitoring the performance of these systems and predicting their output. One of the methods of analyzing the data is by cluster analysis or clustering. Kaufman and Rousseeuw (1990) describe cluster analysis as the "art of finding groups in data". More formally, the aim of cluster analysis is to identify groups of similar objects, where objects in a cluster are more similar to each other than objects in different clusters (Halkidi et al., 2001). This technique can therefore reveal patterns that may exist and that may have not yet been identified. One of the applications of clustering is classification. Classification is a technique used in this thesis for understanding and characterizing the solar irradiance patterns in Durban. Some previous studies that have applied clustering to irradiance data are discussed below.

4.2 Review of irradiance clustering

As mentioned in Chapter 1, there have been previous studies that have investigated classification of days based on solar irradiance profiles. Among these classification studies, the most commonly used classification parameter was the clearness index, k_t. In addition, the studies focused mainly on using the classification results for optimization and sizing of solar power plants (such as PV) and analyzing their performances.

A study by Zagouras et al. (2013) used clustering for identification of geographical zones of GHI, and suggested that these zones could be used in forecasting. However, this is different from the present study which is concerned with clustering daily irradiance profiles for the purpose of forecasting at a single geographical location, namely Durban. Zagouras et al. (2014) also used clustering for determining coherent climatic zones of GHI, where the knowledge of the coherent zones could be used for deciding on the appropriate placement of solar farms.

For South Africa, Zhandire (2017) proposed a "solar utility index" (SUI) for solar resource classification that uses clustering to produce classes of SUI for each of nine radiometric stations across the country, including Durban. This differs from the clustering presented here in that the SUI is a daily average of beam and diffuse horizontal irradiance, whereas the present work considers diurnal variation of beam irradiance. In addition, the primary aim of the study was to provide a new solar resource index for classification in South Africa, and did not aim exclusively to characterize the solar irradiance patterns in Durban or investigate the possibility of forecasting using classification results. None of the above-mentioned studies used the clustering results for solar forecasting.

As mentioned earlier, studies that have used clustering and classification for forecasting include those by McCandless et al. (2014) and McCandless et al. (2015). In common with the present work, cloud regimes were identified by *k*-means clustering (McCandless et al., 2014), but by contrast McCandless et al. (2014) tested short-term forecasts, up to 3 hours ahead, rather than day forecasts. In another study, Benmouiza and Cheknane (2013) used clustering combined with ANNs to generate forecasts for hourly global radiation.

The aim of this thesis was to use clustering for forecasting. Of particular interest is the work of Badosa et al. (2013), Badosa et al. (2015) and Jeanty et al. (2013) that serves as a basis for much of the present study. A characterization of mesoscale and local-scale solar irradiance variability for Reunion Island was conducted by Badosa et al. (2013). They combined satellite and ground-based measurements to analyze the variability in cloudiness and surface irradiance from diurnal to

seasonal scales. Since the largest amount of irradiance variability occurs at the diurnal scale, the authors applied clustering analysis to three irradiance parameters to characterize this daily variability. Results showed that the island's diurnal variation can be classified into five main irradiance regimes. Badosa et al. (2015), also explored the variability in local solar irradiance conditions and cloud cover dynamics of Reunion Island using irradiance regimes presented in Badosa et al. (2013) together with synoptic wind and relative humidity parameters. The novelty of this study was the use of only exogenous variables to make day-ahead solar irradiance predictions.

As mentioned earlier, Jeanty et al. (2013) also investigated irradiance patterns for Reunion Island. The study used cluster analysis applied to daily profiles of direct horizontal irradiance fraction, k_b, defined in Chapter 2. Similarly to Badosa et al. (2013), the study by Jeanty et al. (2013) revealed five dominant patterns for the island. However, only four of the five patterns, corresponding to Clear, Cloudy, AM Clear and PM Clear conditions, are positively correlated with the classes obtained by Badosa et al. (2013).

In the present study, in a similar manner to Jeanty et al. (2013), minute-resolution irradiance profiles were pre-processed by Principal Component Analysis (PCA) and then clustered by the hierarchical and *k*-means methods. To illustrate the PCA and clustering methods, clustering of minute-resolution values of k_b was used as an example. Although *k*-means will be used for classification and forecasting, a comparison with hierarchical clustering is also presented. Clustering of minute-resolution k_b using hierarchical and *k*-means clustering serves as a comparison between the two methods. As described in Chapter 2, for a period of one year in Durban, daily profiles of *D* and *G* were recorded at one-minute intervals between 8:30-16:30. Minute-resolution profiles of k_b for the year were then derived using equation (2.16). Each minute of k_b is a point in a 481-dimensional space. An example of a k_b profile for Durban is shown in Figure 4.1. Processing of the data and implementation of the methods were done in MATLAB using the Statistics Toolbox.

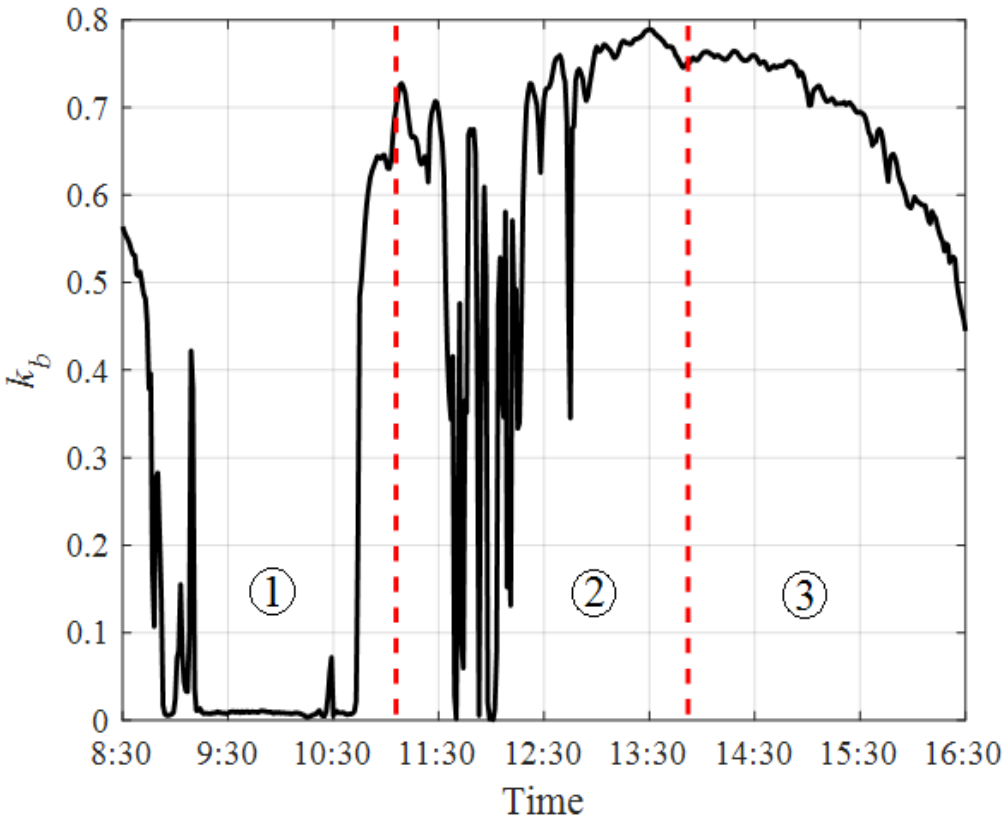


Figure 4.1: Typical k_b profile for Durban on 21 April 2017. The regions indicated by ①, ② and ③ respectively denote morning, midday and afternoon periods of the day during this interval.

4.3 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a common pre-processing method used for dimension reduction and was applied for this purpose in the present study. Tufféry (2011) describes PCA as a technique used for projecting a cloud of individual points on to subspaces with fewer dimensions. For a set of p variables describing n individuals, each individual can be represented by a point in a p-dimensional space R_p. The set of individuals is said to be a "cloud of points".

As described by Jolliffe (2002), the main idea of PCA is to reduce the dimensionality of a data set in which there are a large number of variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of axes, called Principal Components (PCs), which are ordered successively so that the first few components retain most of the variance present in the original data. The first Principal Component (or first new axis) accounts for the highest variance in the data, the second Principal Component the second highest variance

and so on. Therefore, a small number of Principal Components may be able to account for most of the variation in the data with minimal loss of information. Prior to applying PCA, the original data should be normalized. In this case, normalization refers to re-scaling of the data so that variables of two different scales may be compared. In general, the purpose of normalization is so that in the case of applying PCA to variables with different scales, the variable with the largest scale does not dominant over all variables in the data set. For the clustering presented in this thesis, B_n and D_n were normalized to the CSM.

To illustrate the PCA technique, Figure 4.2 (a) shows the technique being applied to a 3-dimensional data set. The first, second and third dimensions are the morning, midday and afternoon averages of k_b, respectively, and correspond to the regions indicated by ①, ② and ③ in Figure 4.1. The first dimension, PC1, accounts for most of the variance in the data, and the data is most spread out along the direction of this component. The second dimension, PC2, also accounts for a considerable amount of variance, but less than PC1. The spread of the data along this direction is significant but smaller than the spread along the first. The third dimension, PC3, accounts for only a small amount of variance i.e. very little spread along this direction. Therefore, only the first two components are retained, meaning the data can be reduced to 2-dimensions since most of the information about the original data is contained within them. As shown below in Figure 4.2 (b), PC1 and PC2 become the new set of axes onto which the data points are transformed.

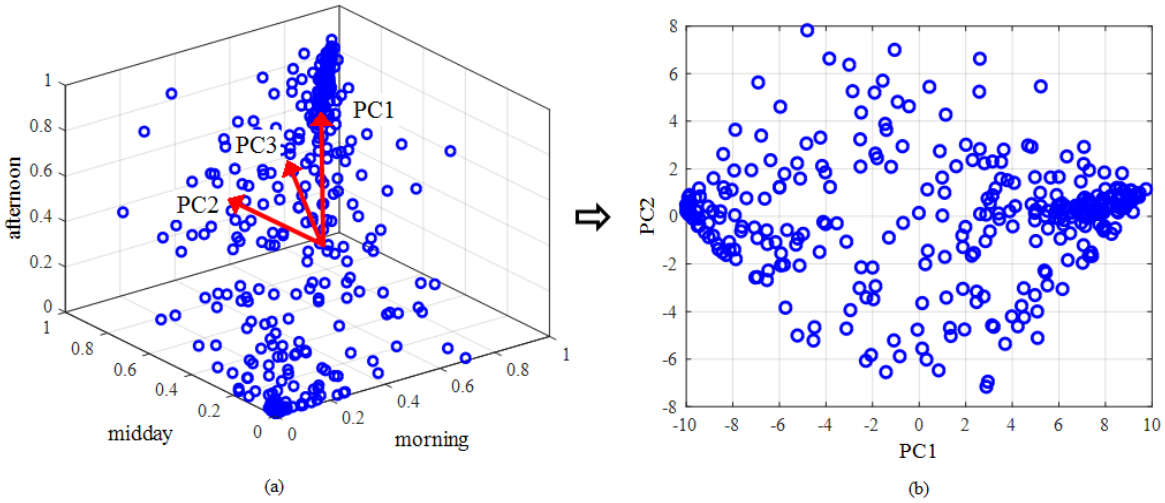


Figure 4.2: Example of how PCA is applied to (a) 3-dimensional cloud of points to reduce it to (b) 2-dimensions. The data has the highest variance along the first dimension, PC1, and the second highest variance along the second dimension, PC2. The third dimension, PC3, accounts for only a small amount of variance, and therefore can be eliminated.

4.4 Choice of Principal Components

The primary objective of PCA is to reduce data with a large number of dimensions to fewer dimensions. This technique is useful since there are a large number of dimensions from minute-resolution irradiance observations throughout the day. Each Principal Component (or axis) accounts for a certain percentage of the total variance, and the aim is to find the minimum number of Principal Components that explain most of the variance in the data. In the case of data with a large number of dimensions the first two Principal Components may not be sufficient to describe the information in the data and so it is often the case that more than two components are included. In order to choose the fewest number of Principal components to be retained, the cumulative percentage of the total variation is used, where the minimum number of components with a cumulative percentage of 90% is retained (Jolliffe, 2002). The variance of each Principal Component is computed from the elements in that component. The percentage variance, p_n, of the *n*-th component is computed by

$$p_n = \frac{var_n}{var_{TOT}}, \tag{4.1}$$

where var_n and var_{TOT} are the variance of the *n*-th Principal Component and the total variance of all components, respectively.

PCA was applied to the 481-dimensional set of minute-resolution k_b profiles to determine the fewest number of Principal Components that explain at least 90% of the data variance. Figure 4.3 shows a *scree* plot that is used to view the Principal Components and their fraction of the total variance arranged in descending order. The scree plot shown in Figure 4.3 is of the first 10 Principal Components. It can be seen that the first component has a high variance contribution, indicating that this component itself contains 74% of the information about the data. Each of the subsequent components explain less than 10% of the variance. For the case of minute-resolution k_b, analysis of the cumulative percentage showed that the first 5 Principal Components account for 90% of the variance, and hence were retained.

4.5 Comparison between clustering methods

For this thesis two clustering methods, hierarchical and *k*-means, were applied to daily minute-resolution k_b profiles for Durban. Application of clustering to the same set of k_b profiles serves as a comparison between the two methods. Even though only *k*-means clustering was used for

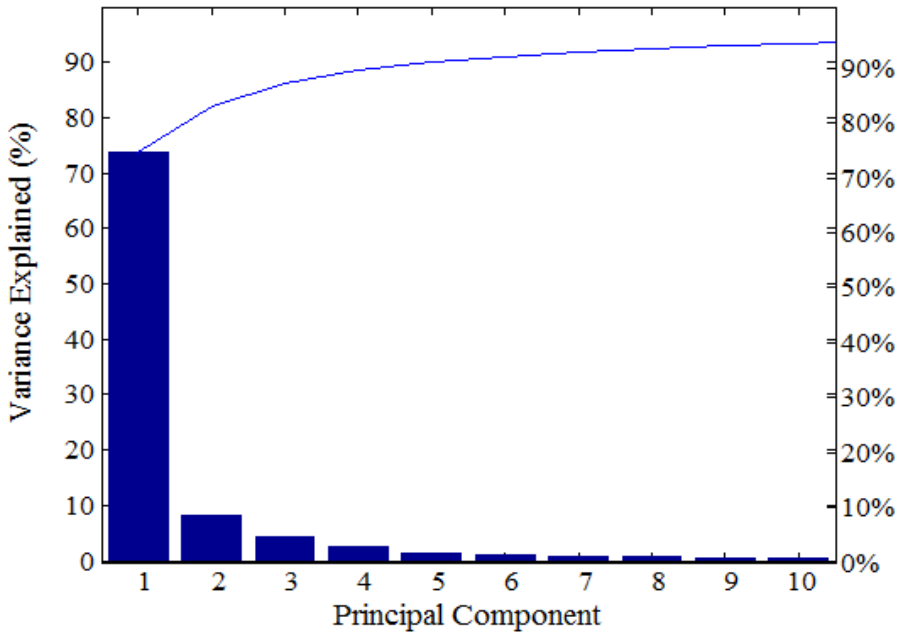


Figure 4.3: Scree plot which shows that the first 10 components that account for a total of 95% of the variance in the data. The first 5 Principal Components account for 90% of the variance. The first component alone accounts for 74% of the variance and the second component accounts for 8%.

classification and forecasting, it was interesting to observe how hierarchical clustering performs i.e. if the hierarchical method can produce similar clustering to *k*-means. In the sections below, hierarchical clustering is first presented and thereafter *k*-means clustering. The clusters resulting from both methods are compared using a Silhouette Index (SI). Furthermore, they can be combined as a hybrid method, where the number of clusters obtained from the hierarchical method is used to initialize *k* in the *k*-means method. However, for the reasons outlined at the end of this chapter, the present work used only *k*-means and applied the silhouette criterion to choose the optimal number of clusters.

4.6 Optimal cluster number and cluster validation

In order to quantify the compactness and separation of each cluster the SI (Silhouette Index) was used as a guide. A *silhouette* is a graphical display of how well an object has been clustered. Comparison of cluster compactness and separation is displayed by combining the silhouette of every object into a single plot. The silhouette shows which objects lie well within their cluster, and which

ones are close to the border of clusters (Rousseeuw, 1957). The SI for each point provides a value based on the relation of that point to all the points in its respective cluster compared to points in other clusters. According to Rousseeuw (1957), let $a(i)$ be the average distance between object *i* and all the other members of its own cluster. For another cluster C, let $d(i, C)$ be the average distance between object *i* and the members of C. Let $b(i)$ be the minimum of $d(i, C)$ over all the other clusters. Then SI is given by

$$SI = \frac{b(i) - a(i)}{max\{b(i)a(i)\}}. \tag{4.2}$$

The SI value ranges from -1 to 1. An SI that approaches unity is indicative of a object belonging to a coherent cluster and one can say that the object has been "well-clustered". Alternatively, an SI value approaching -1 indicates that the object is not well-suited to that cluster (Rousseeuw, 1957). The average SI for an individual cluster (denoted here as $\overline{SI}_C$ where C is a cluster label), or for the entire set of objects (denoted here as $\overline{SI}_{TOT}$) is used as an index of overall clustering compactness. Lletí et al. (2004) consider a SI of 0.6 to be a good clustering result, which will be used in this thesis as a criterion for acceptable compactness. In practice it may often be difficult to achieve this value for all clusters. Therefore, a compromise between the SI and the number of clusters (i.e. the largest number of clusters that give the highest $\overline{SI}_{TOT}$) is sought, and used to determine the best clustering solution for the data set (Benmouiza and Cheknane, 2013). A detailed discussion on silhouette interpretation and validity can be found in Rousseeuw (1957). The silhouette plots are used to analyze cluster compactness for both hierarchical and *k*-means methods.

4.7 Hierarchical clustering

A Hierarchical clustering procedure is one which successively merges smaller clusters into larger ones (agglomerative), or divides larger clusters into smaller ones (divisive). This process may be represented by a tree-like structure called a *dendrogram* which depicts the relationship between objects or clusters. The dendrogram shows how single objects and clusters are grouped together at each step and provides a measure of similarity between them. This similarity is the Euclidean distance where if the distance between two clusters is small then they are close together and hence more similar. If the distance is large then the clusters are less similar. The Euclidean distance on

the y-axis on the dendrogram is the distance between the singletons, and thereafter are the distances between centroids of clusters.

Figure 4.4 demonstrates the hierarchical clustering method by way of an example. Figure 4.4 (a) shows a set of 5 points of morning and afternoon averages of k_b. The method starts off by assuming each point is a cluster on its own. Then the clusters that are closer together merge to form a new cluster. The distance between clusters 3 and 4 is 0.04, and are clearly closer to each other than to other clusters, hence they merge to form cluster 6. The distances between clusters 2 and 1 and 2 and 5 are 0.23 and 0.21, respectively. Therefore, clusters 2 and 5 merge to form cluster 7. For merging clusters, Ward linkage was used. There are also other linkage options such as single, average and complete. However, according to Tufféry (2011) the Ward linkage (Ward, 1963) is considered the most effective linkage method. At the last step, clusters 6 and 8 merge.

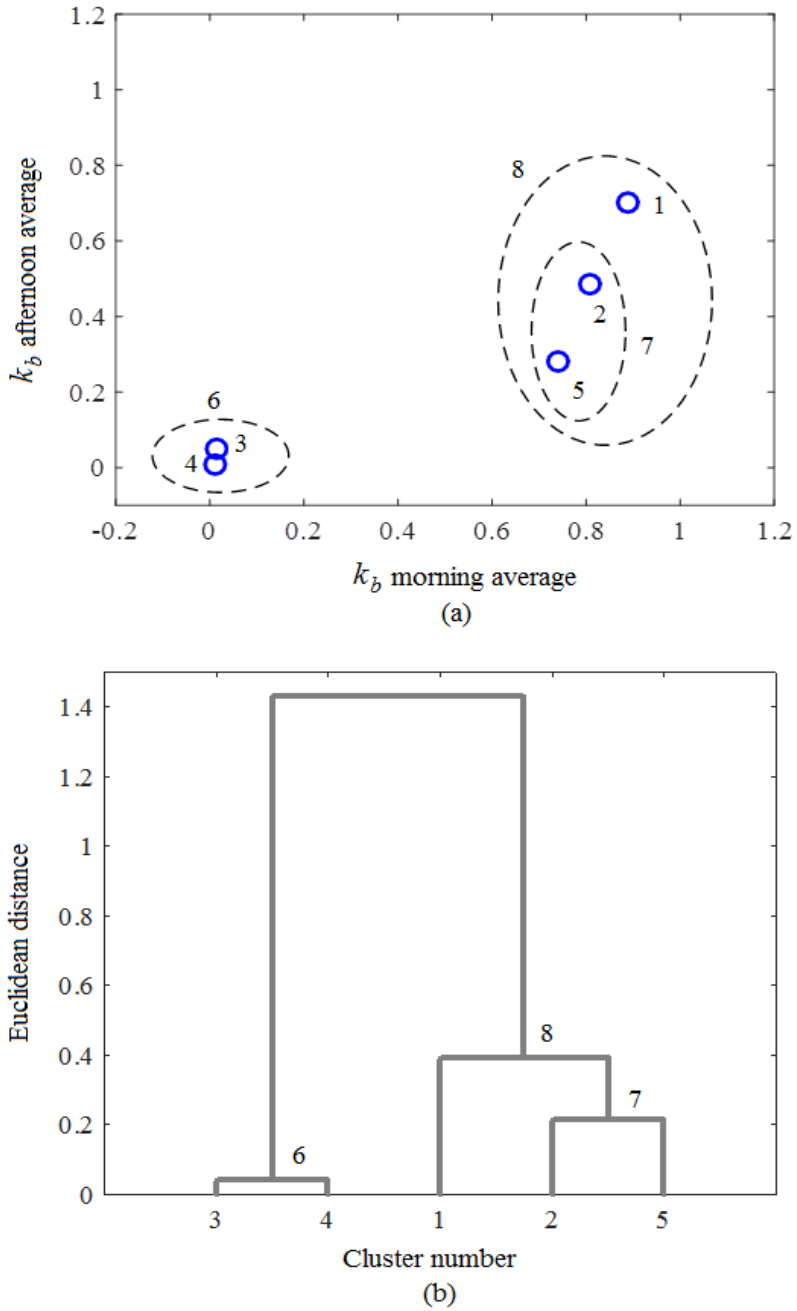


Figure 4.4: (a) Five points that will be clustered using the hierarchical method. Each point starts off as a cluster on its own. (b) Dendrogram showing how clusters in (a) were merged. Clusters 3 and 4 and 2 and 5 were merged at distance 0.04 and 0.21, respectively. The centroid of cluster 7 was merged with cluster 1 at a distance of 0.4. Lastly, the centroids of clusters 6 and 8 were merged at distance 1.4.

To demonstrate the use of hierarchical clustering on the minute-resolution k_b profiles, the method was applied to the k_b Principal Components. The Ward's linkage method was used with the Euclidean distance as the metric. According to equation 4.3, the Ward's linkage method minimizes the total within-cluster sum of the squared error (SSE) when merging two clusters. The Ward's distance between two clusters *A* and *B* having centers *a* and *b* and frequencies n_A and n_B, is given by

$$d(A, B) = \frac{d(a, b)^2}{{n_A}^{-1} + {n_B}^{-1}}, \tag{4.3}$$

where *a* and *b* are the centroids of clusters *A* and *B*, respectively. Once all of the objects are clustered the dendrogram is produced. Cutting the dendrogram at a desired level will result in a set of disjoint groups (or clusters). However, in the present study, the optimal number of clusters was not known a priori. The choice of the optimal number of clusters in order to specify the level at which the dendrogram should be cut must be decided using an appropriate method. The present work used the cluster sum of squares as a guide to finding the level at which the dendrogram should be cut to yield the optimal number of clusters.

Computing the cluster sum of squares for different clustering solutions, can be used as a guide for choosing the optimal number of clusters. According to Tufféry (2011), the total sum of squares, *I*, of the cluster is the weighted mean of the squares of the distances of the individual points from the cluster center (or centroid), and is given by

$$I = \sum_{i \in I} p_i (x_i - \bar{x})^2, \tag{4.4}$$

where $\bar{x}$ is the mean of x_i and p_i is the weight associated with observation i. In a similar manner, the sum of squares of a cluster is computed with respect to its own center

$$I_j = \sum_{i \in I_j} p_i (x_i - \bar{x_j})^2. \tag{4.5}$$

If the data is partitioned into *k* clusters, each with sums of squares I_1, . . ., I_k, then within-cluster sum of squares, I_W, is

$$I_W = \sum_{j=1}^{k} I_j. \tag{4.6}$$

The between-cluster sum of squares, I_B, is defined as the mean of the squares of the distances of the centers of each cluster from the global center, given by

$$I_B = \sum_{j \in clusters} \left(\sum_{i \in I_j} p_i \right) (x_j - \bar{x})^2. \tag{4.7}$$

Therefore, the total sum of squares is the sum of the within-sum of squares and between-sum of squares, given as

$$I = I_W + I_B. \tag{4.8}$$

The illustration in Figure 4.5 depicts the total sum of squares for a set of points which is the sum of the within-sum of squares and between-sum of squares.

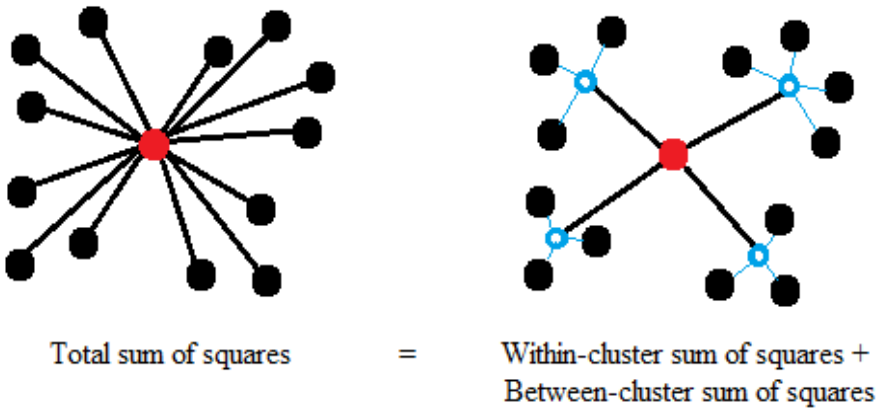


Figure 4.5: The total cluster sum of squares (I) is the sum of the within-sum of squares (I_W) and between-sum of squares (I_B). Global cluster centers are indicated in red. Adapted from Tufféry (2011).

The value for I_W can be used to find the optimal number of clusters present in the data. If all points belong to one cluster i.e. k = 1, I_W will be high since there will be points that are far away from the cluster centroid, thus increasing the sum of squares. As k increases, I_W decreases since there are more centroids and the clusters become more homogeneous. However, finding the largest k is not necessarily the best clustering solution. Instead the number of clusters should be increased such that if the last significant decrease in I_W occurs when moving from k to k + 1 clusters, the partition into k + 1 clusters is correct. This is demonstrated in Figure 4.6.

To decide on the level of cutting of the dendrogram and to obtain the k_b clusters, Figure 4.6

shows I_W computed for values of k ranging from 1 to 10. The curve starts off at a high value for $k = 1$ which is expected since all objects are assigned to one cluster. As k increases, I_W decreases dramatically and thereafter begins to flatten out as k approaches 10. Tufféry (2011) recommends that the value of k should be chosen such that on moving from k to $k + 1$, there is an insignificant decrease in I_W. However, Tufféry (2011) provides no criteria for what constitutes an insignificant decrease of I_W, so choosing the cut-off value of k is a matter of judgement. For the minute-resolution k_b data, the last significant decrease was chosen to be $k = 3$ to $k = 4$. Therefore, the optimal number of clusters is set to 4. The dendrogram can now be cut at the level that yields 4 clusters.

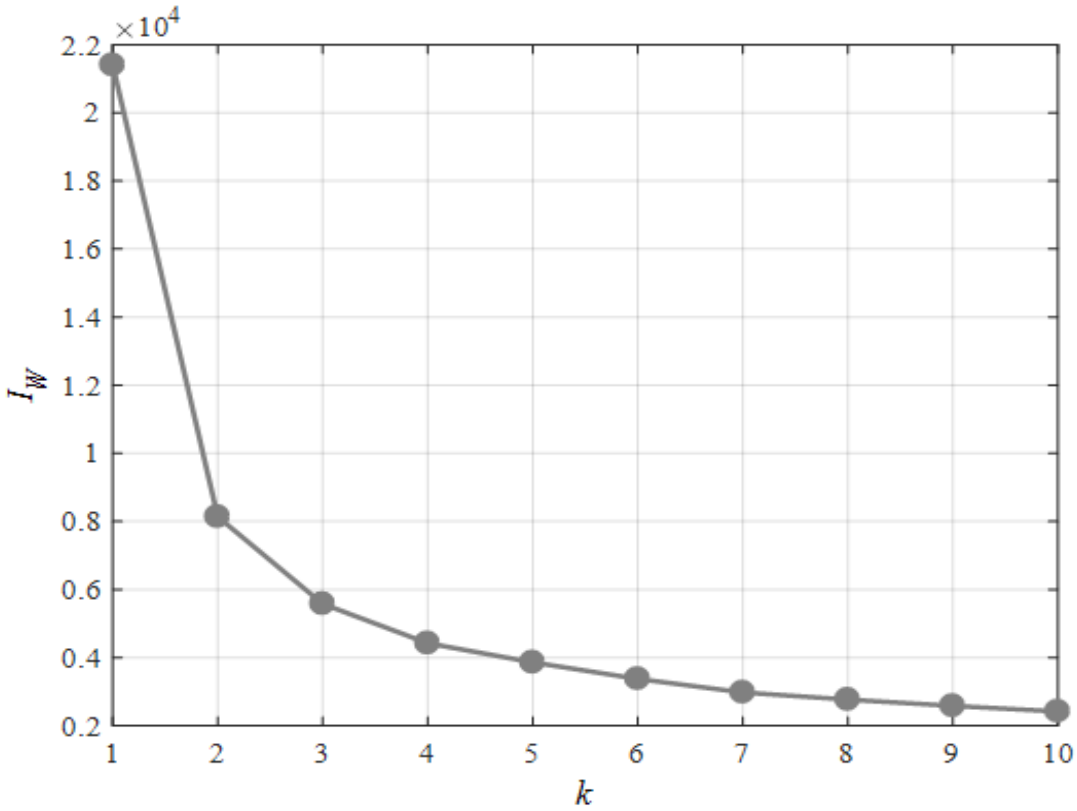


Figure 4.6: Within-cluster sum of squares for varying values of k, for k_b clusters using the hierarchical method. For $k = 1$, I_W is high. As k increases, I_W decreases dramatically and thereafter begins to flatten as k approaches 10. The optimal value of k is 4 since moving from $k = 3$ to $k = 4$ results in a small decrease in I_W.

The silhouette plot for the hierarchical k_b clusters of the Durban data is given in Figure 4.7. Cluster 1 has a low $\overline{SI}_C$ and is rather weakly clustered. Cluster 2 also has a low $\overline{SI}_C$. The $\overline{SI}_C$ for Cluster 3 is above 0.8 indicating a compact cluster. Cluster 4 has a slightly lower $\overline{SI}_C$ than Cluster

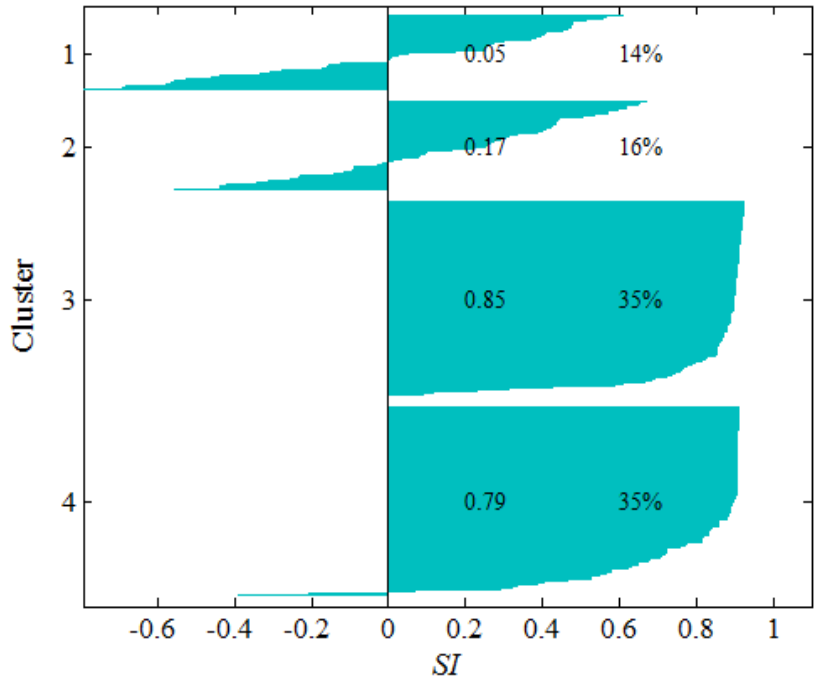


Figure 4.7: Silhouette plot for clusters 1 to 4. Clusters 1 and 2 have low $\overline{SI}_C$ indicating less compact clusters. Clusters 3 and 4 have high $\overline{SI}_C$ indicating compact clusters. The percentage of days in each cluster is also given. Negative SI values are days that lie closer to the border of the cluster.

3, but nevertheless still sufficiently high to be regarded as compact. The percentage of days in each cluster is also given. Days with negative are close to the border of two clusters and comprise 11% of days. For all 4 clusters produced by the Ward's hierarchical method, the $\overline{SI}_{TOT}$ was found to be 0.61.

The Ward's hierarchical clustering procedure applied to the PCA-reduced k_b data, produced the dendrogram in Figure 4.8. Using the with-cluster sum criterion in Figure 4.6, the dendrogram was cut at the level that produced 4 clusters. A cluster map showing the first two Principal Components is given in Figure 4.9. Cluster 3 and Cluster 4 are relatively compact. However, Clusters 1 and 2 are less compact.

4.8 Partitional clustering: *k*-means

The *k*-means algorithm, developed by MacQueen (1967), is a well-known and commonly used partitional clustering algorithm. It belongs to the family of clustering algorithms which require the a priori specification of a desired number of clusters. Given a set of *n* objects, this method constructs *k* partitions of the data, where each partition represents a cluster and where $k \leq n$. The partition divides the data into *k* groups such that each group contains at least one object. The primary aim of *k*-means clustering is to optimize the following objective function:

$$E_D = \sum_{i=1}^{c} \sum_{x \in C_i} d(x, m_i), \tag{4.9}$$

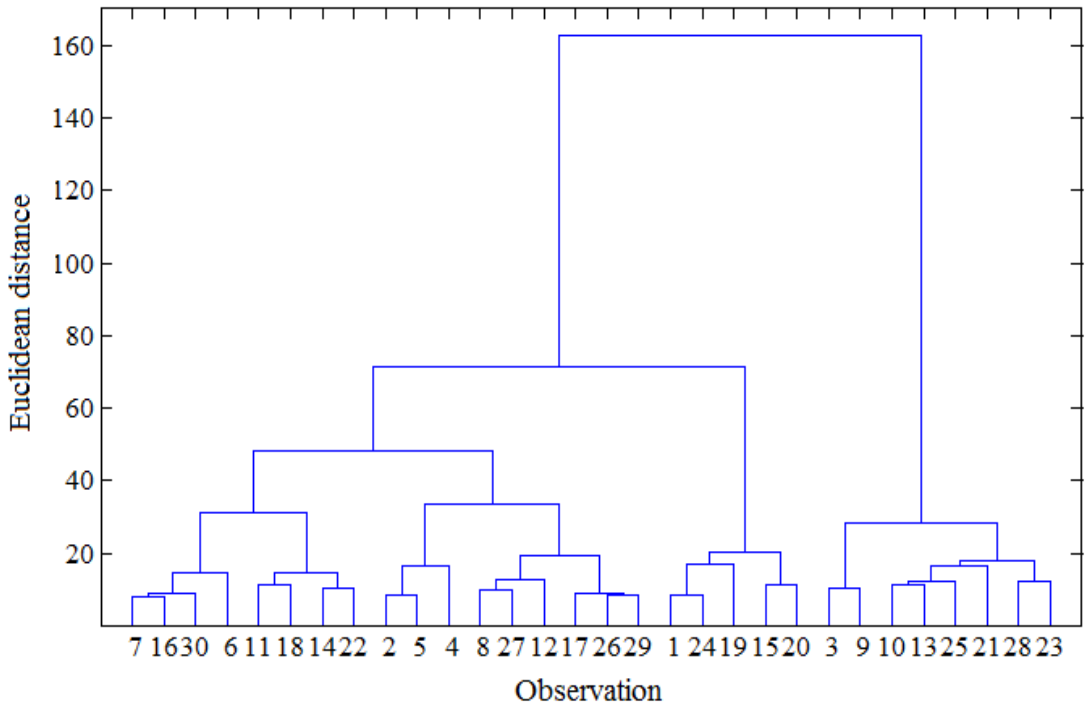


Figure 4.8: Dendrogram depicting the clustering in the data, from single observations to one final cluster. The Euclidean distance scale is a measure of how similar two observations and clusters are. The lower the horizontal line that links two observations or clusters, the smaller the Euclidean distance between them and the more similar they are. Large vertical gaps on the dendrogram indicate observations or clusters are less similar to each other.

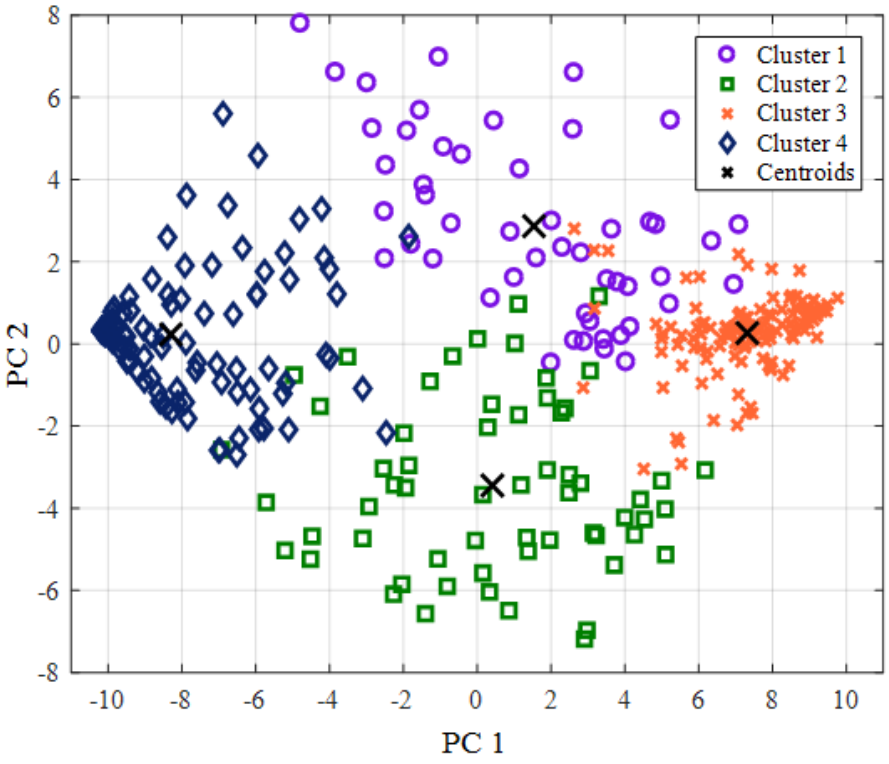


Figure 4.9: Cluster map of k_b PC1 and PC2 that shows how days are clustered using the hierarchical method. PC1 accounts for 74% of the variance and PC2 accounts for 8%. Cluster 3 and Cluster 4 have a high degree of compactness (i.e. high $\overline{SI}_C$) and Clusters 1 and 2 have a low degree of compactness (i.e. low $\overline{SI}_C$).

where E_D is the criterion function, m_i is the center of the cluster C_i and $d(x, m_i)$ is the Euclidean distance between a point x and m_i. The criterion function attempts to minimize the distance of each point from the center of the cluster to which the point belongs (Halkidi et al., 2001). In general, the k-means iterative clustering method is implemented as follows:

1. Choose a k value.
2. Select k objects arbitrarily. Use these as the initial set of k centroids.
3. Assign each of the objects to the cluster for which it is nearest to the centroid.
4. Re-calculate the centroids of the k clusters, which is done by averaging the members of the cluster.
5. Repeat steps 3 and 4 until the centroids no longer move (Bramer, 2007).

There is no guarantee that *k*-means finds the global minimum, but it does find a local minimum for a given initial choice of centroids. In order to check for variation in clustering due to different initial centroids, *k*-means was run several times.

The *k*-means clustering was applied to the resulting data, after the PCA was applied. The best clustering solution was chosen by considering the results for various values of *k*, guided by $\overline{SI}_{TOT}$ as well as values of $\overline{SI}_C$ for each cluster. The aim was to find the "natural" clustering, which corresponds with finding the largest number *k* of compact clusters. The guideline used was to maximize *k* while keeping $\overline{SI}_{TOT}$ greater than 0.6, and also seeking to maintain $\overline{SI}_C$ as high as possible for each of the clusters obtained. Figure 4.10 shows how $\overline{SI}_{TOT}$ varies with *k*, where *k* ranges from 2 to 10. The largest $\overline{SI}_{TOT}$ is for *k* = 4, which is the optimal clustering solution for k_b using *k*-means. Although *k* = 2 has a distinctly higher $\overline{SI}_{TOT}$, as mentioned, the aim was to find the largest *k* with an $\overline{SI}_{TOT}$ greater than 0.6.

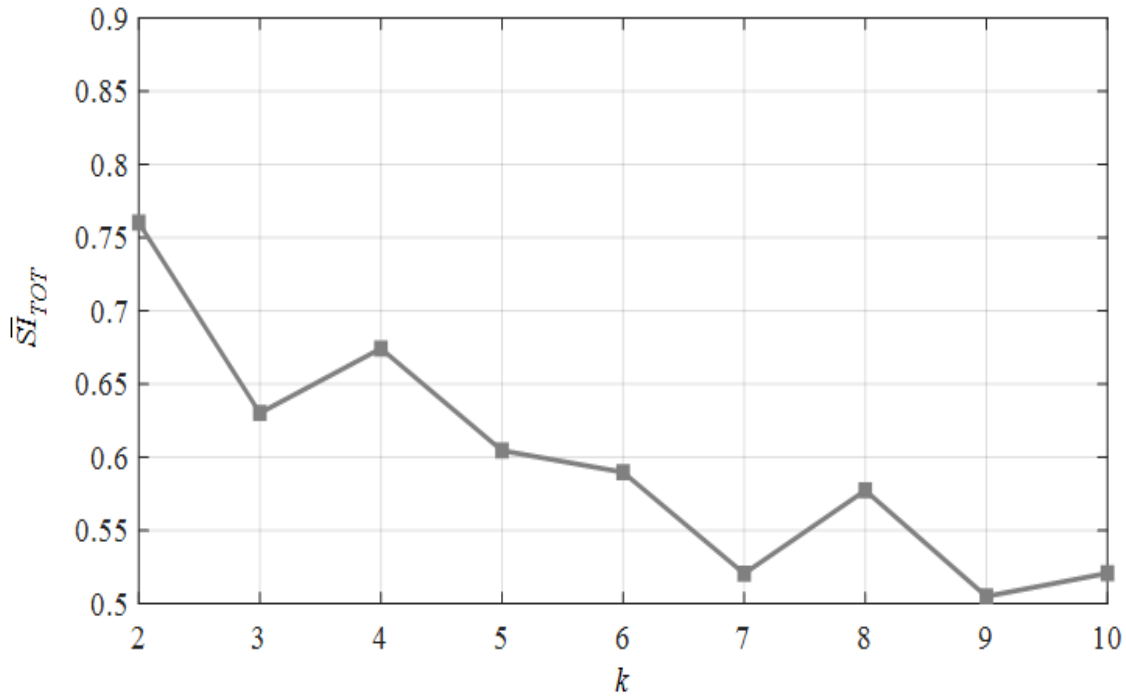


Figure 4.10: $\overline{SI}_{TOT}$ for varying values of *k* for k_b where values above 0.6 correspond to *k* = 2, 3, 4.

The cluster map depicting the *k*-means clusters is given in Figure 4.11. The variation of the data is highest along the *x*-axis (PC 1) and second highest along the *y*-axis (PC 2). The silhouette plot, Figure 4.12, shows how well the k_b data has been clustered. For Clusters 1 and 2 there are large SI_C values indicating that these days belong in these clusters. Clusters 3 and 4 have lower SI_C values

which suggests that days in these clusters are closer to the border of the clusters. A total of 3% of days for the entire data set have negative SI values. This indicates that having 4 clusters is a good clustering solution for the data since 97% of the days fall into a cluster. The $\overline{SI}_{TOT}$ for the entire data set is 0.74, which again shows that the data has been well-clustered.

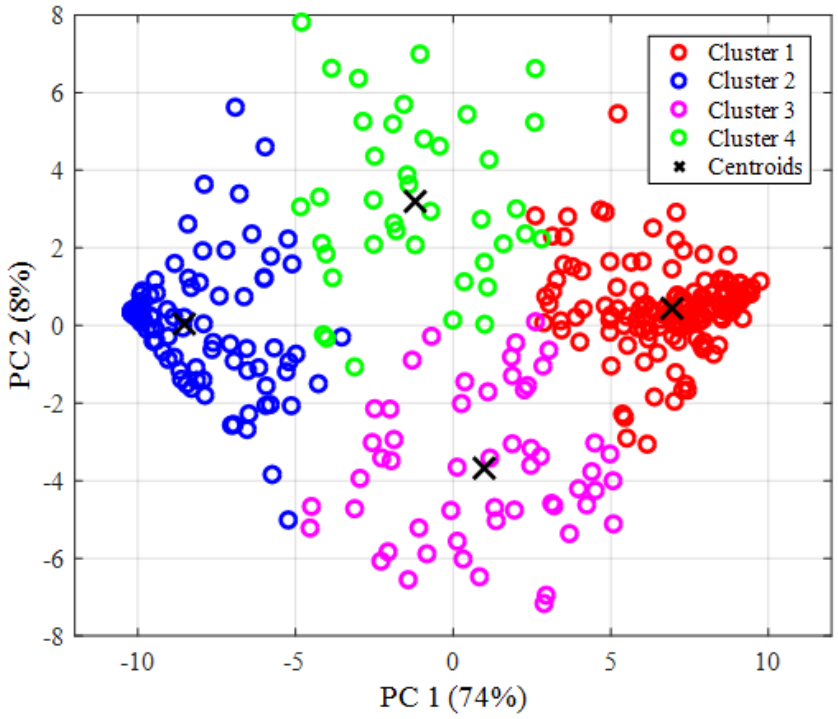


Figure 4.11: Cluster map of k_b PC1 and PC2 that show how days are clustered using the *k*-means method. Clusters 1 and 2 are compact while Clusters 3 and 4 are significantly less compact. A total of 97% of the days fall into a cluster.

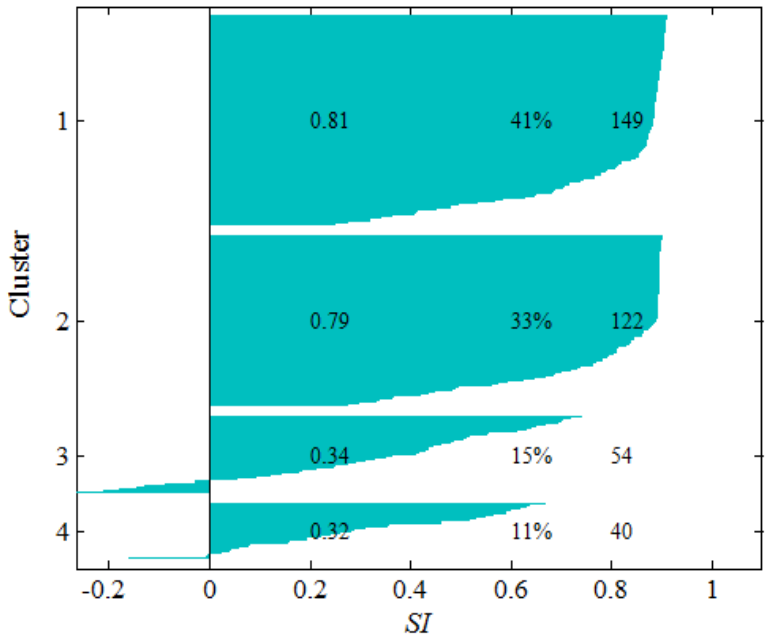


Figure 4.12: $\overline{SI}_C$ for each of the 4 clusters and the corresponding frequency of day in each cluster. Clusters 1 and 2 have $\overline{SI}_C > 0.75$, which indicates compact clusters. Clusters 3 and 4 have $\overline{SI}_C$ below 0.35 and hence are not compact. A total of 3% of days for the entire data set have negative SI values.

Two clustering techniques were applied to one year of minute-resolution k_b profiles. Due to the large number of dimensions i.e. 481, PCA was applied as a pre-processing step prior to clustering. The first technique, Ward's hierarchical clustering, produced 4 clusters as the optimal clustering solution. The choice of the number of clusters was based on analysis of the within-cluster sum of squares for varying *k* from 1 to 10. Ward's clustering produced two dominant clusters each with $\overline{SI}_C$ above 0.7, and two weak clusters with values for $\overline{SI}_C$ less that 0.2.

The second technique, *k*-means clustering, also produced 4 clusters as the optimal solution as guided by the $\overline{SI}_{TOT}$. As shown in Figure 4.10, only values of *k* = 2, 3 and 4 have $\overline{SI}_{TOT}$ above 0.6. $\overline{SI}_{TOT}$ for *k* = 5 is 0.6 and $\overline{SI}_{TOT}$ is less than 0.6 for the remaining *k* values. Therefore, choosing the highest *k* from with $\overline{SI}_{TOT}$ values above 0.6 gives *k* = 4 as the best solution. The silhouette plot for the *k*-means clusters show that Clusters 1 and 2 are compact clusters. Clusters 3 and 4 are less compact with $\overline{SI}_C$ of 0.34 and 0.32, respectively.

As mentioned earlier, it is often difficult to attain $\overline{SI}_C > 0.6$ for all clusters. Both clustering techniques produce two compact clusters and two weakly compact clusters. However, the *k*-means clustering produces a slight improvement on the weakly compact clusters, where for both clusters $\overline{SI}_C$ is greater than 0.3. In addition, as compared to hierarchical clustering, the *k*-means method produces a higher $\overline{SI}_{TOT}$, more clusters with higher $\overline{SI}_C$ and fewer days with negative SI values.

Overall, both clustering techniques produce a similar clustering solution i.e. 4 clusters. Out of the clustering sample which comprised one year of daily k_b profiles, 88% of days were found to be in the same class when clustered by both techniques. Therefore, it is sufficient to apply only one of them to find irradiance patterns to be used for classification and forecasting.

An important feature of the *k*-means is that the distance criterion between a point and its cluster centroid, is the same as the error that is computed between individual profiles and the profile of the cluster mean. This error is the smallest for *k*-means because the minimization of point to centroid distances of a cluster is the essence of the *k*-means clustering algorithm. This is however not the case for the hierarchical clustering. Furthermore, the *k*-means clustering is a widely used method as compared to hierarchical clustering. Therefore, for this thesis, clustering of all variables will be done using the *k*-means clustering method.

Chapter 5

Classification of irradiance profiles

This chapter presents a classification of irradiance profiles where the profiles are either normalized beam and diffuse irradiance, variability in the beam irradiance, or a combination of normalized beam and diffuse irradiance. As shown in the previous chapter clustering using the hierarchical method yields similar patterns to clustering by *k*-means. Therefore, as mentioned in Chapter 4, clustering of all profiles are done by the *k*-means method.

5.1 Clustering of profiles

For this thesis, clustering of several radiometric variables was investigated. For classification, minute-resolution profiles of B_n were clustered, and the classes established have distinct diurnal profiles that characterize the irradiance patterns in Durban. To match the temporal resolution of the NWP cloud cover output, hourly-resolution $\bar{B}_n$ profiles were also clustered. This further led to the investigation of clustering of V_B, the variability in B_n, to regain information that could have been lost through averaging. Since B_n was the main clustering variable for distinguishing sky conditions, and since V_B may contain information about B_n, this then led to the clustering of the combination of B_n and V_B. Furthermore, the combination of B_n and D_n was also considered. Classification of daily profiles is based on the clusters that will be determined, so the terms "cluster" and "class" are used interchangeably.

As shown at the end of Chapter 4, clustering was applied to minute-resolution k_b. However, a significant characteristic of k_b is that it is dependent on seasonality. To illustrate this, k_b profiles for winter and summer solstices for Durban are shown in Figure 5.1. These are theoretical k_b profiles

that were derived using equation 2.16 and the D and G components of the Ineichen clear sky model (Ineichen and Perez, 2002), discussed in Chapter 2. On a clear day, k_b rises in the first hour after dawn to near unity and declines in late afternoon during the last hour before sunset. As the seasons change, sunrise and sunset times change accordingly so the k_b profiles vary in the start and end times of the rising and declining phases. This seasonal variation in profile, even for clear days, creates differences that are not due to cloud conditions. These differences can be eliminated by using a normalized quantity such as B_n.

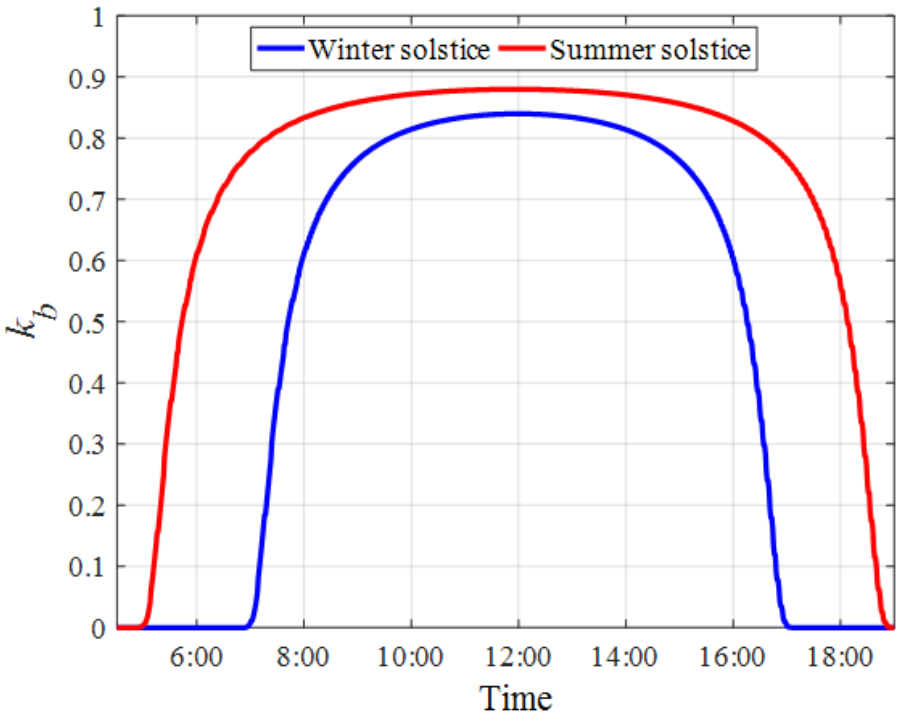


Figure 5.1: Theoretical k_b profiles using clear sky components of D and G for winter and summer solstices in Durban.

Therefore, due to the seasonal dependency of k_b causing differences between profiles not due to cloud conditions, B_n was used for classification and forecasting. Shown in Figure 5.2, are the normalized profiles for B_n for winter and summer solstices. B_n is zero before sunrise and after sunset and these times vary with season. However, within the interval 8:30-16:30 i.e. the time during which clustering was applied, the shape of the B_n profile is independent of seasonality. Between the sunrise and sunset times its shape depends on cloud conditions but for a clear day its profile is a horizontal line. The upper limit of B_n is close to 1, depending on the atmospheric

turbidity. Since this study focuses on clustering irradiance profiles based on how they are affected by clouds, introducing the B_n variable was an appropriate step in the method, as will be shown in the present chapter. In a similar manner to B_n, D_n was also introduced. The value of D_n exceeds 1 depending on the amount of cloud cover. Knowledge of the beam irradiance component is required for CST plants, but both beam and diffuse components are required for PV plants. Furthermore, if the two components are known then the normalized global irradiance, G_n, can be obtained.

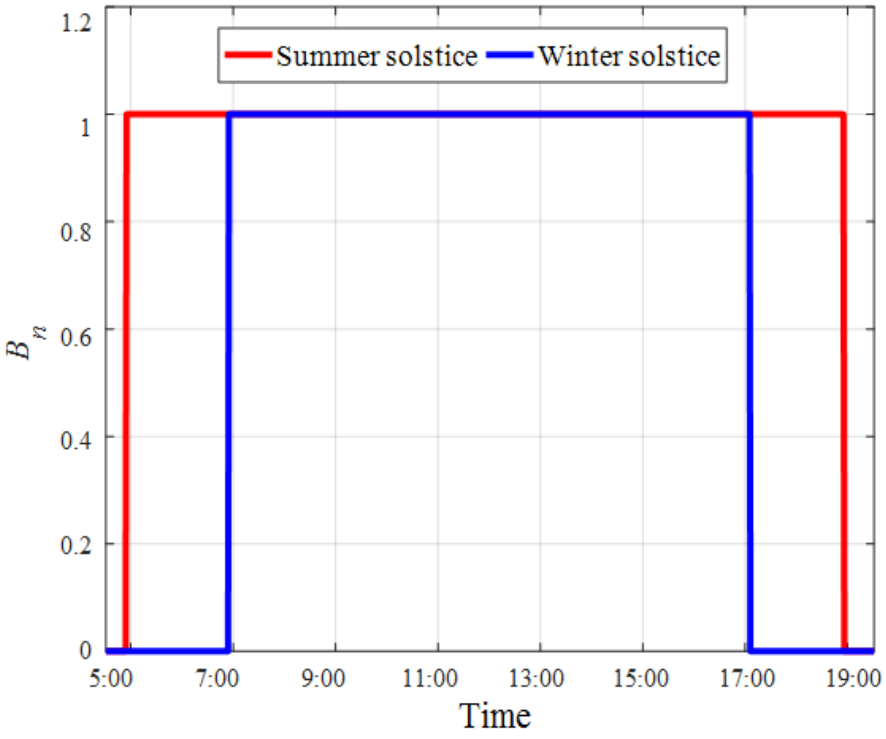


Figure 5.2: Theoretical clear sky B_n profiles for winter and summer solstices in Durban. Within 8:30-16:30, the shape of the B_n profile is independent of seasonality.

5.2 Minute-resolution normalized beam irradiance, B_n

Each daily profile of B_n consisted of a set of 481 values for each minute during the interval 8:30 to 16:30 (solar time). Therefore, each B_n profile corresponds to a point in a 481-dimensional space, which was reduced to a low-dimensional space by PCA. The first 8 components accounted for 90% of the variance, using equation 4.1, and were hence retained. Figure 5.3 shows a scree plot, where the percentage variance of the first 10 Principal Components are shown. The first component alone

accounts for 72% of the variance and the second component accounts for 8%. The cumulative percentage variance of all 10 components is 91%.

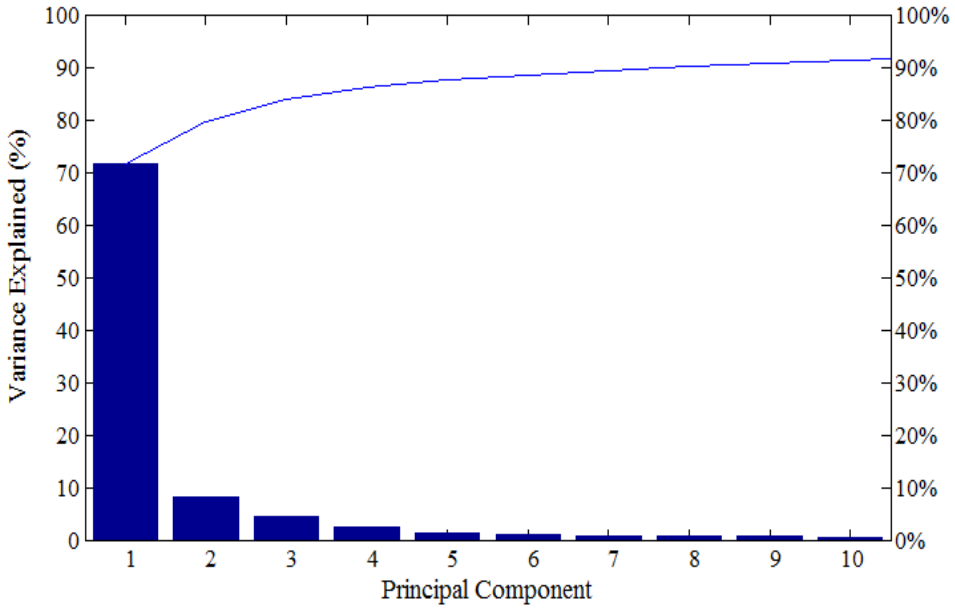


Figure 5.3: The first 10 Principal Components of B_n and the percentage variance explained. The cumulative percentage for the first 8 and 10 Principal Components are 90% and 91%, respectively.

The *k*-means algorithm was applied to the PCA-reduced set of B_n profiles. The *k*-means cluster map of the first two Principal Components is given in Figure 5.4. The *x*-axis (PC1) accounts for 72% of the variance and the *y*-axis (PC2) accounts for 8%. We refer to Cluster 1, 2, 3 and 4 as Class A, B, C and D, respectively. The frequency of days in each cluster and the cluster $\overline{SI}_C$ value are presented in Table 5.1. Classes A and B have $\overline{SI}_C > 0.7$, but $\overline{SI}_C$ for Classes C and D are much lower. The total number of days with $SI < 0$ is 19 and $\overline{SI}_{TOT}$ = 0.63.

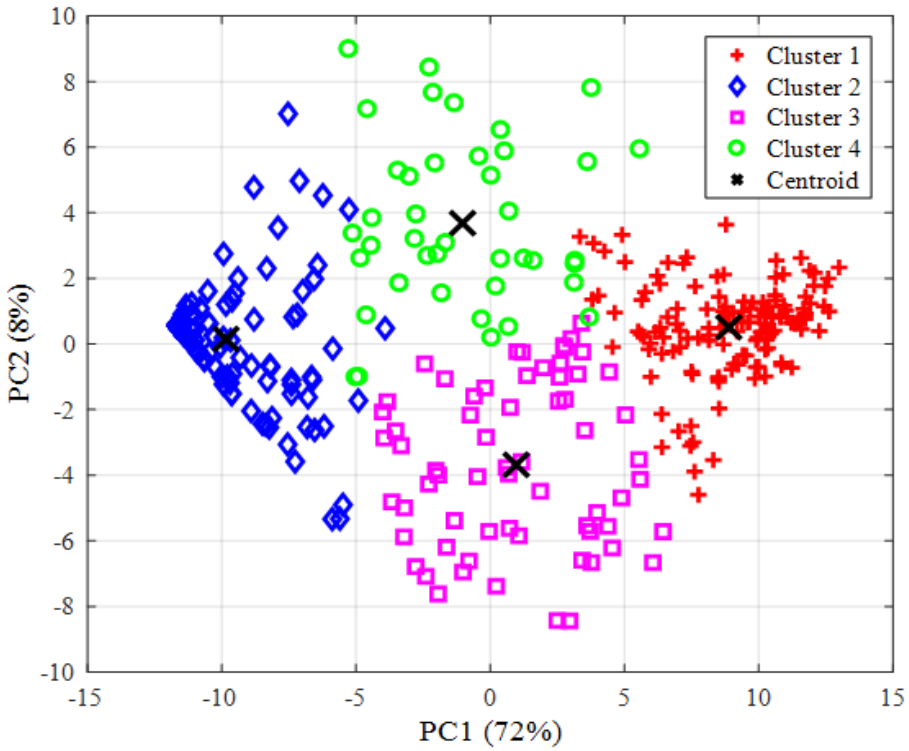


Figure 5.4: Cluster map of the first two Principal Components (PC1 and PC2) of the minute-resolution B_n profiles.

Table 5.1: Summary of B_n clustering.

Class	Cluster	Frequency of days	Proportion	$\overline{SI}_C$	$\overline{SI}_C < 0$
A	1	135	37%	0.78	0
B	2	124	34%	0.78	0
C	3	65	18%	0.31	10
D	4	41	11%	0.21	9

5.3 Physical interpretation of the classes

To better understand the physical meaning of the classes established through clustering, class mean profiles are shown in Figure 5.5 and 5.6. The profiles describe the diurnal patterns of each class. Figure 5.5 suggests that there are four classes of diurnal variation in beam irradiance on the half-day scale, namely sunny all day (Class A), cloudy all day (Class B), sunny in the morning and cloudy in the afternoon (Class C) and cloudy in the morning and sunny in the afternoon (Class D). These classes are hence identified as irradiance pattern classes, which are labeled as follows: Class A: sunny, Class B: cloudy, Class C: sunny AM-cloudy PM and Class D: cloudy AM-sunny PM.

From the mean profiles in Figure 5.5, Class A has B_n values > 0.8 throughout the day, therefore these days are sunny with little or no cloud cover. Class B has low B_n levels throughout the day i.e. B_n values that are < 0.2. This indicates that these days experience cloudy and overcast conditions. Class C is one which is characterized by relatively high B_n values (> 0.6) in the morning up until about midday and then decreases during the rest of the day. However, even though the mornings are regarding as sunny, they are not as sunny as the morning of day that belongs to Class A. This could indicate that days in this class have some cloud present that lowers the B_n levels during the sunny period, and these clouds become more dominant over clear sky conditions during the afternoon. Days in Class D start off cloudy in the morning and become sunny ($0.4 < B_n < 0.65$) in the afternoon. Similar to Class C, the B_n levels during the sunny region in Class D do not reach the same maximum values in the afternoon, as compared to days in Class A.

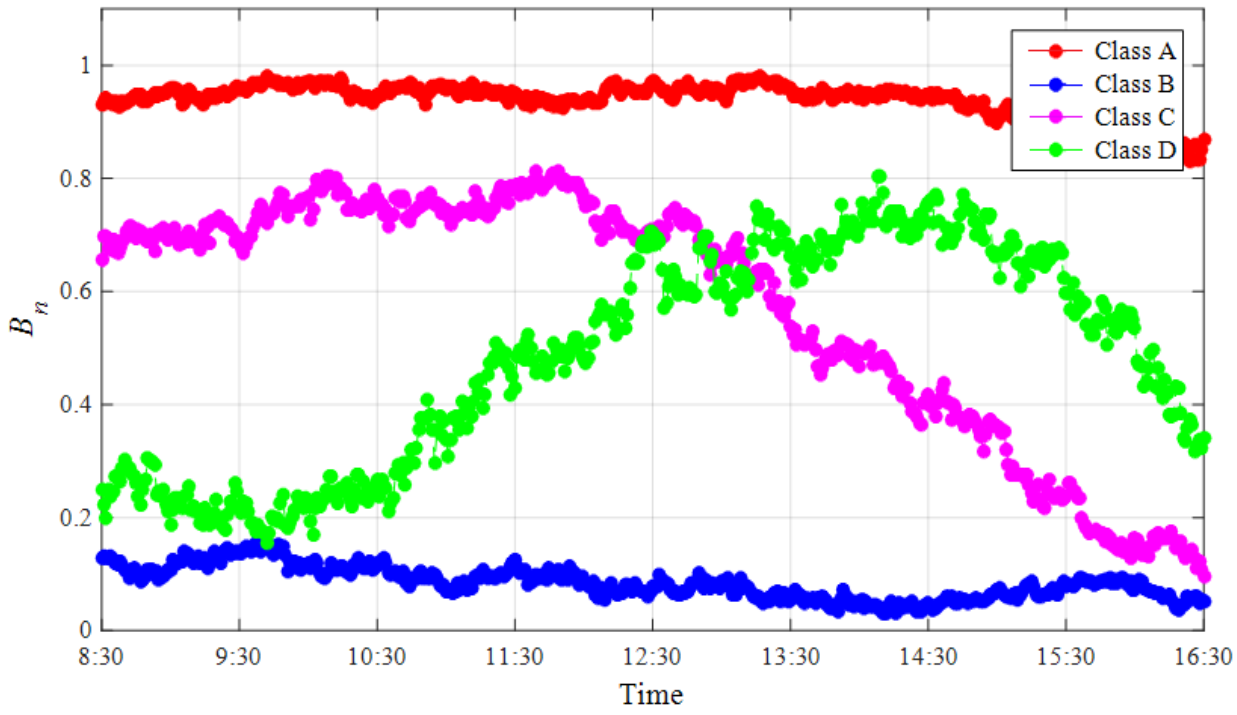


Figure 5.5: Mean profiles of the B_n classes. Cloudy and sunny conditions are characterized by low and high B_n values, respectively. Class A has high B_n values throughout the day, while Class B has low B_n values. Class C has high B_n levels in the morning and low B_n in the afternoon. Class D has low B_n in the morning and high B_n in the afternoon.

The set of associated D_n profiles for each B_n class is given in Figure 5.6. For Class A, D_n is low throughout the day due to the absence of clouds and persistence of clear sky conditions. Class B has high levels of D_n for most of the day due to thick cloud cover. Class C has low levels of D_n in the morning and increases as cloud cover increases toward the afternoon. Class D has the opposite trend of Class C. The low regions of D_n in Class C and D are not as low as Class A indicating that there are some clouds present that increase the diffuse irradiance during these times.

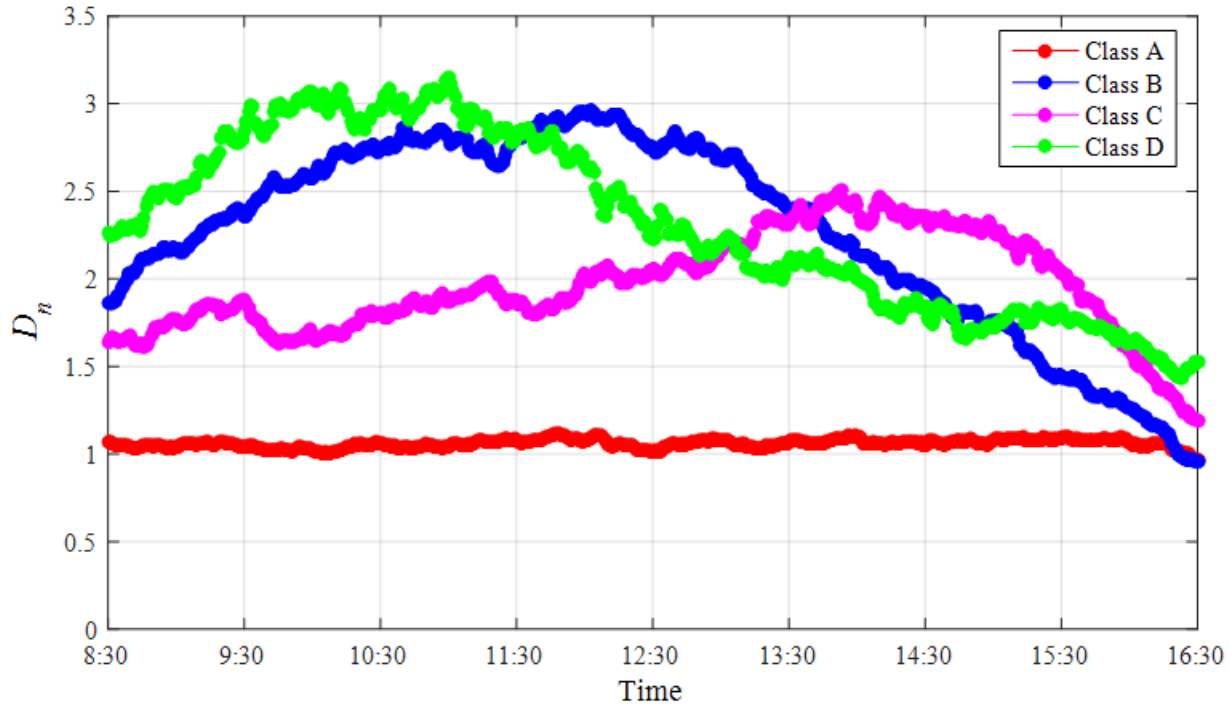


Figure 5.6: Mean profiles of the D_n classes. Cloudy and sunny conditions are characterized by high and low D_n values, respectively. Class A has low cloud levels and low D_n all day. Class B has high D_n for most of the day due to high cloud levels. Class C has low cloud levels in the morning and hence low D_n values and higher D_n in the afternoon. Class D has high D_n in the morning and low D_n in afternoon the afternoon, indicating high levels of cloud cover in the morning and low in the afternoon.

To establish the range of uncertainty in the classes, all the profiles belonging to each class of B_n and D_n are presented in Figures 5.7-5.10. To smooth out the minute-to-minute fluctuations and to visualize the diurnal pattern of the members of a class more clearly, a one hour moving average was applied to the B_n and D_n profiles. In general, there is a large variation in the individual class members for both B_n and D_n. For B_n, Classes A and B have a large portion of their days in a fairly small range, but the range within which the days in Class C and D vary is rather broad. For D_n, Class A has the smallest range of variation indicating that for sunny days the diffuse irradiance does not vary significantly from the mean profile. Classes B, C and D have much larger variation in their profiles. Overall, this shows that even though a day may be in the class, and it is best suited to that class according to its SI value, it can be significantly far away from the mean profile of the class. Nevertheless, this is useful for forecasting as it quantifies the range of variation for a typical day

that falls into each of the B_n and D_n classes.

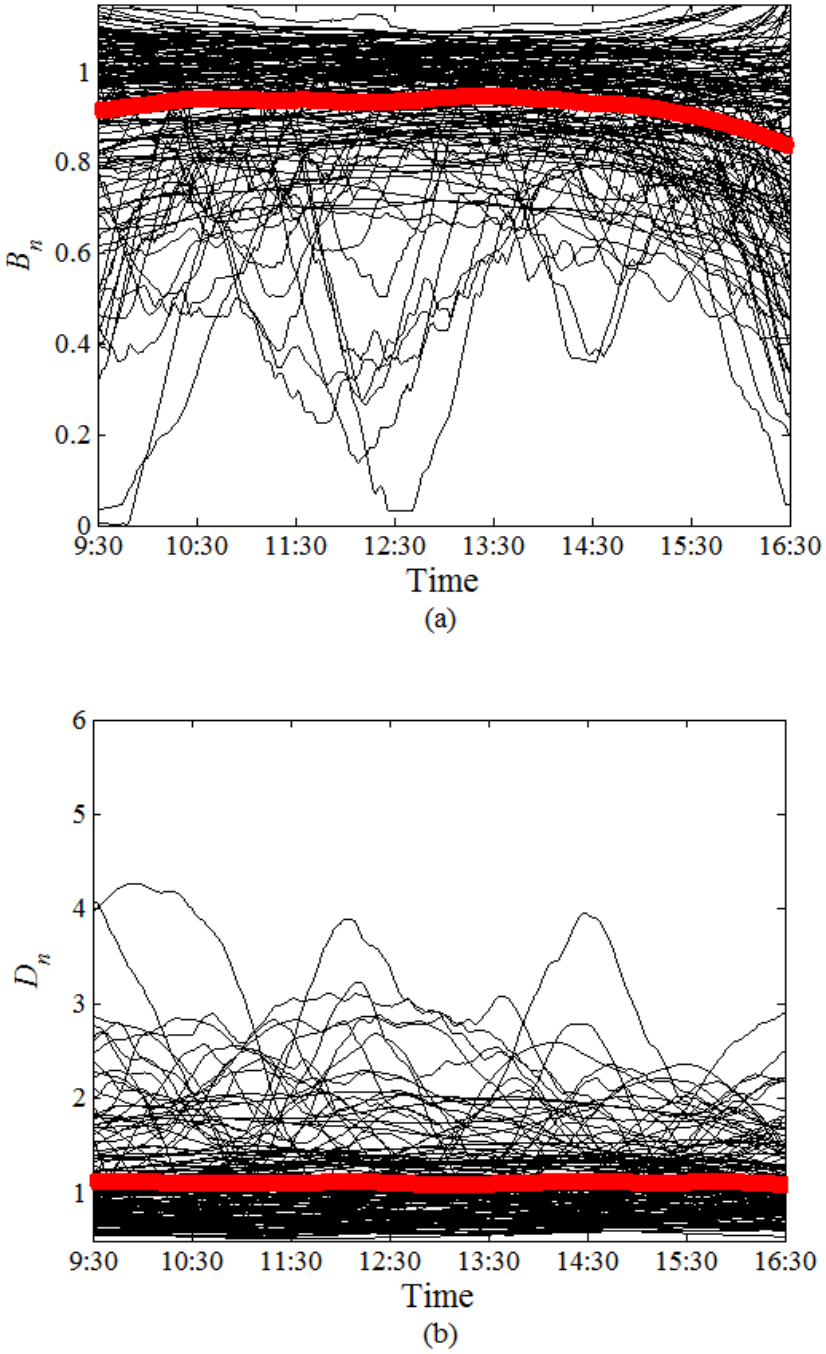


Figure 5.7: Class A minute-resolution profiles after applying a one hour moving average to smooth out minute fluctuations in (a) B_n class members and (b) D_n class members. Also shown is the mean profile of the class superimposed as a thick line.

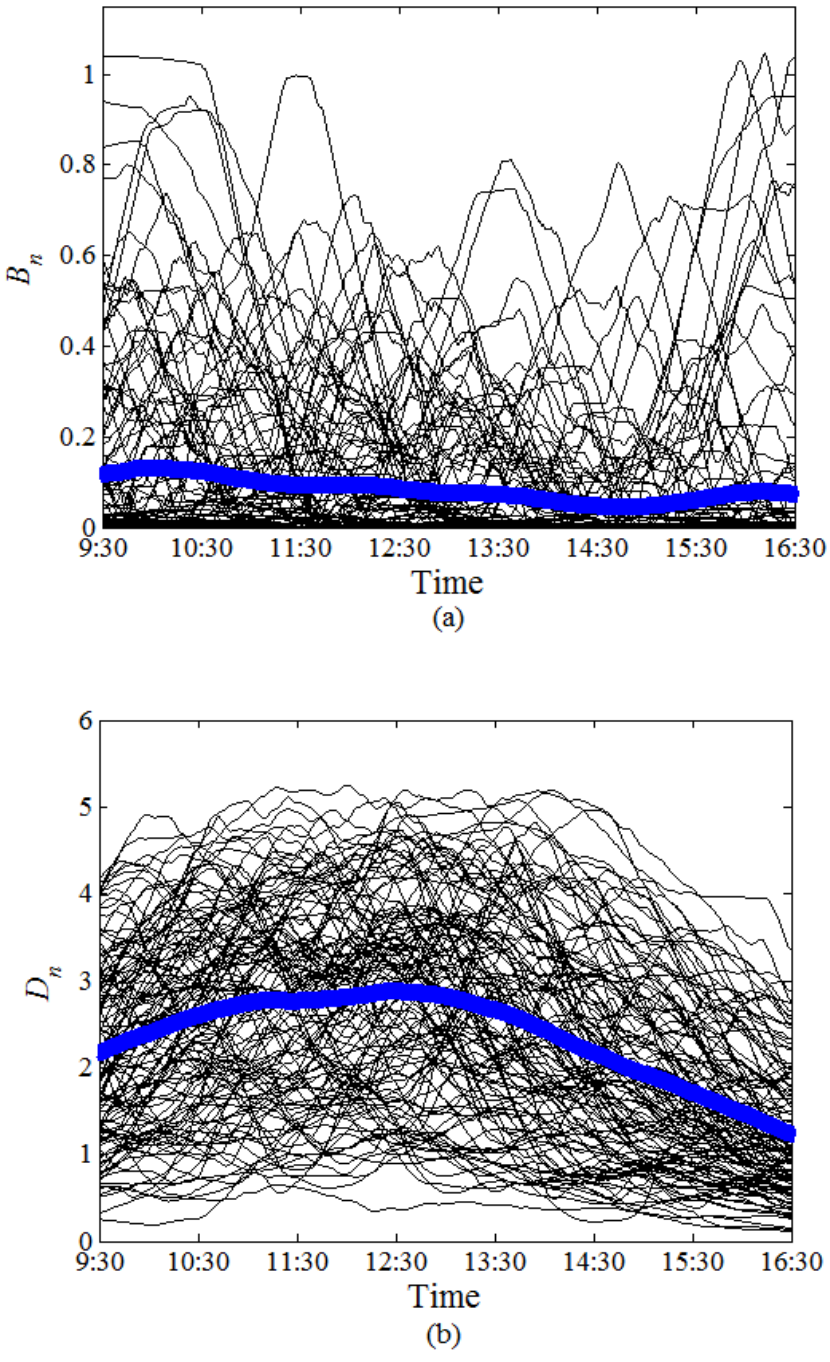
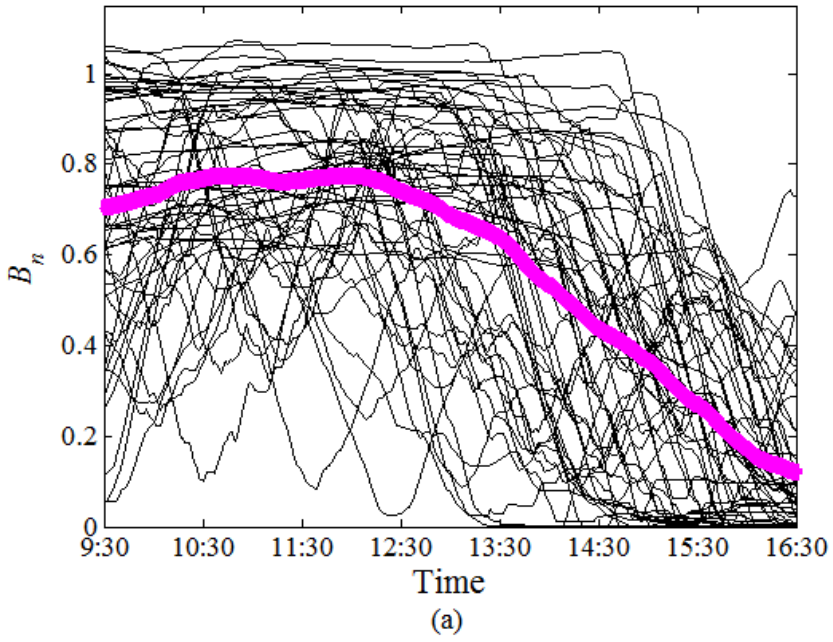


Figure 5.8: Class B minute-resolution profiles after applying a one hour moving average to smooth out minute fluctuations in (a) B_n class members and (b) D_n class members. Also shown is the mean profile of the class superimposed as a thick line.



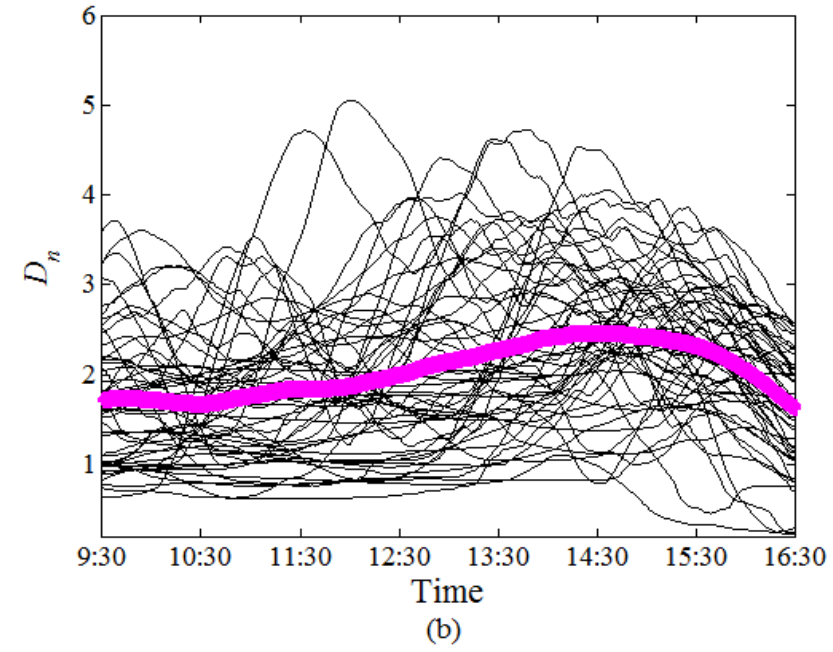
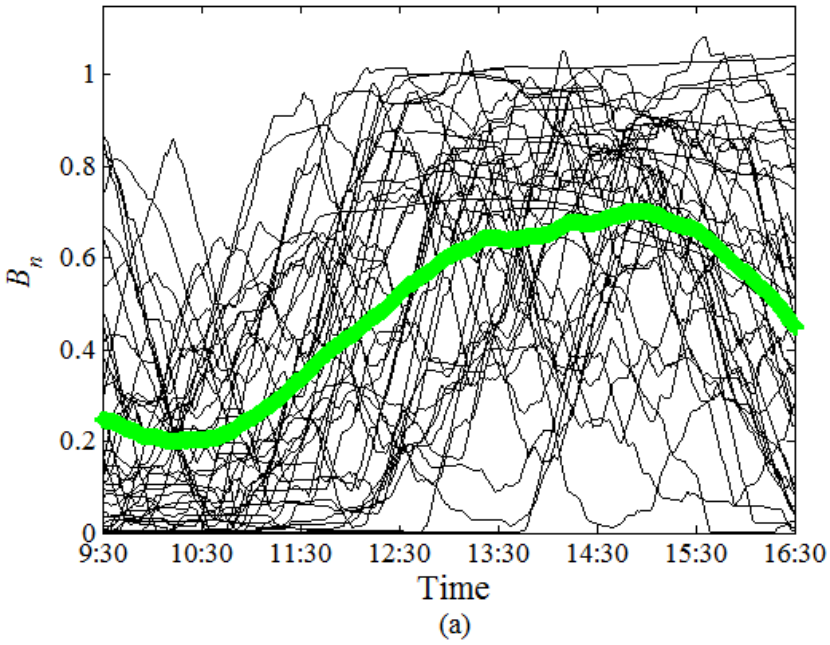


Figure 5.9: Class C minute-resolution profiles after applying a one hour moving average to smooth out minute fluctuations in (a) B_n class members and (b) D_n class members. Also shown is the mean profile of the class superimposed as a thick line.



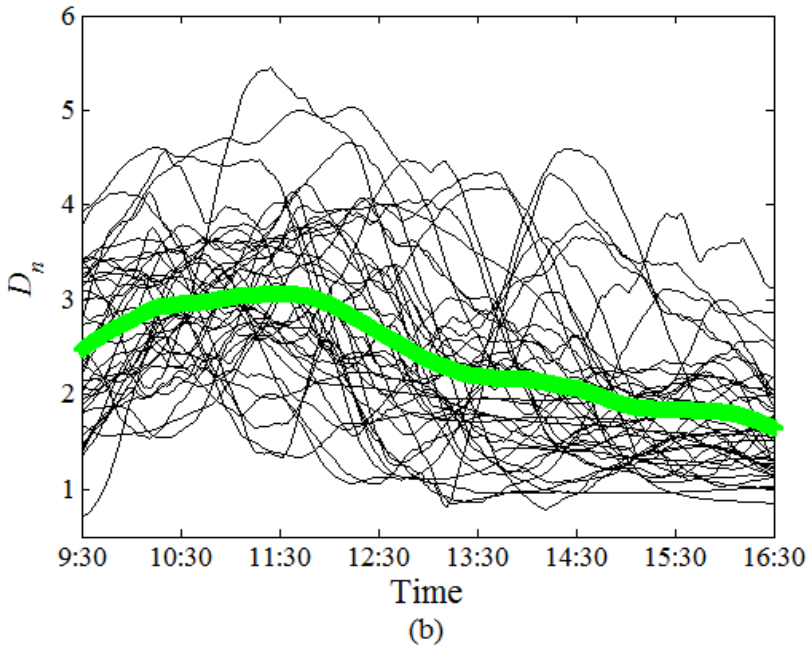


Figure 5.10: Class D minute-resolution profiles after applying a one hour moving average to smooth out minute fluctuations in (a) B_n class members and (b) D_n class members. Also shown is the mean profile of the class superimposed as a thick line.

5.4 Frequency and distribution of the B_n classes

For completeness of characterization of the irradiance patterns in Durban, the annual distribution of B_n classes are presented in Figure 5.11. Class A dominates during April to July. During this period, the months of June and July have 73% and 77% of their days that are sunny. This is consistent with Durban having clearer skies during winter (Zawilska and Brooks, 2011). Class A comprises 10% of the days during the month of November and is therefore the month with lowest proportion of sunny days. Cloudy days that form Class B have highest prevalence during January to March and October to December. These months are characterized by high amounts of opaque cloud cover. In particular, November and December have the highest proportion of Class B days of 63 and 58%, respectively. Class C has the highest occurrence in August and September, with more than 30%, and lowest in June with only 3%. Class D is fairly evenly distributed throughout the year with the exception of July. The months with the largest percentage of Class D days i.e. 20 and 22% are April and October, respectively.

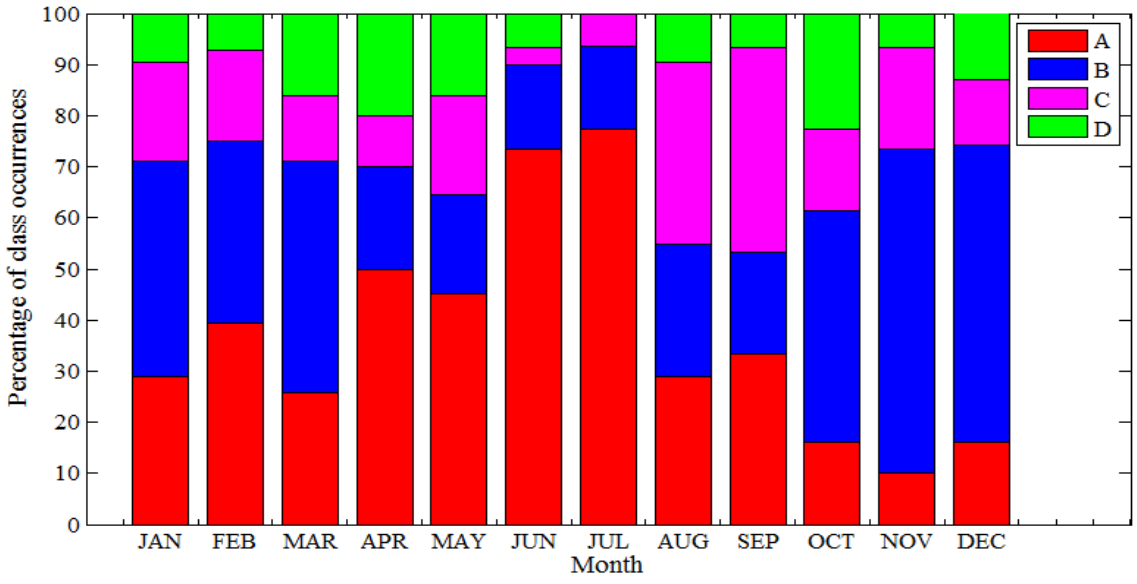
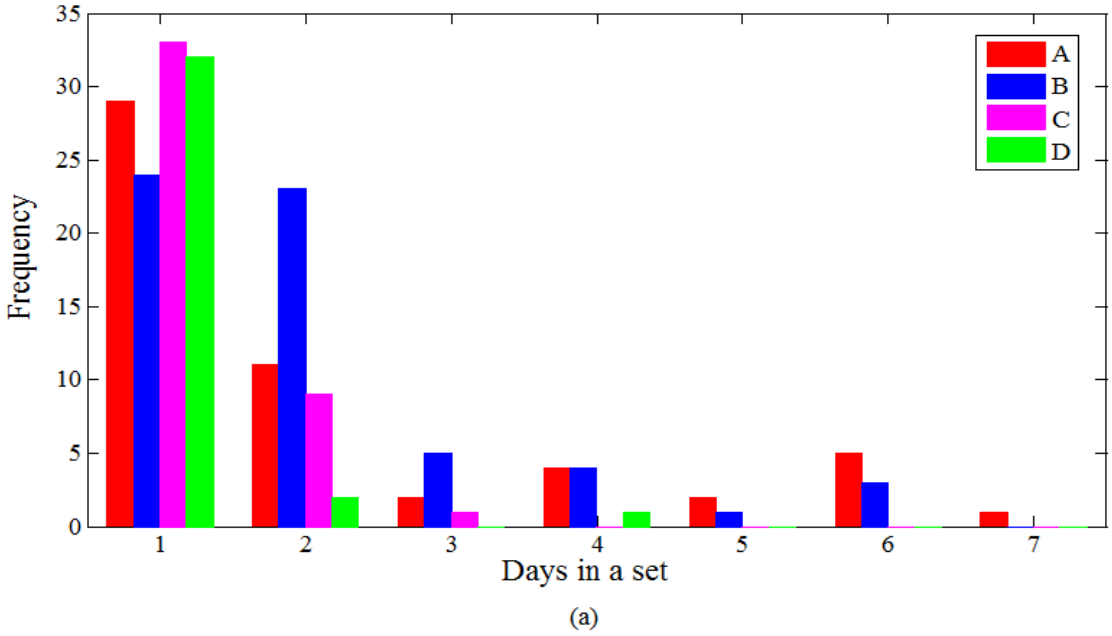


Figure 5.11: Annual distribution of B_n class occurrences. Class A and B dominate during winter and summer, respectively. Class C is most prevalent in August and September. Class D has fairly even distribution throughout the year with the exception of July.

The present work focuses on forecasting the class for the day ahead, at the start of the day. However, some public weather-service providers, such as AccuWeather, provide hourly-resolution

cloud cover forecasts for three days in advance. Therefore, if the current forecasting method is to be extended to forecasting the irradiance class for 2 or more days ahead, how often the classes occur for consecutive days will be of interest. Figure 5.12 (a) shows how frequently classes occur individually and in sets of consecutive days. The days in a set range from 1, which indicates that the class occurred individually, to 7, which indicates that the class occurred in a set of 7 consecutive days. It can be seen that all classes A-D occur most frequently as individual days in the year. Class C has the highest frequency followed by Class D, A and B. Regarding the classes occurring on two consecutive days, Class B is distinctly different from the others by the fact that it has a rather high frequency as compared to classes A, C, and D. In addition, Class B occurs individually in the year almost as often as it occurs in two consecutive days. Interestingly though, only Classes A and B occur in a set of 5 and 6 consecutive days, and only Class A occurs in a set of 7 consecutive days. These frequencies are fairly low.

Figure 5.12 (b) show the frequency of next day class occurrences given the current class of day. More specifically, if a day is currently a Class A, how often is the next day a Class A, or a Class B, C or D. Again, this is useful for forecasting for more than one day ahead. Class A is followed by a Class A day more frequently than any other class. This is also the case for Class B. Class C days however, are most often followed by a Class B, which makes sense since the afternoons of Class C days are cloudy. Class D days have almost similar chance of having the next day be a Class A, B, C or remain a Class D.



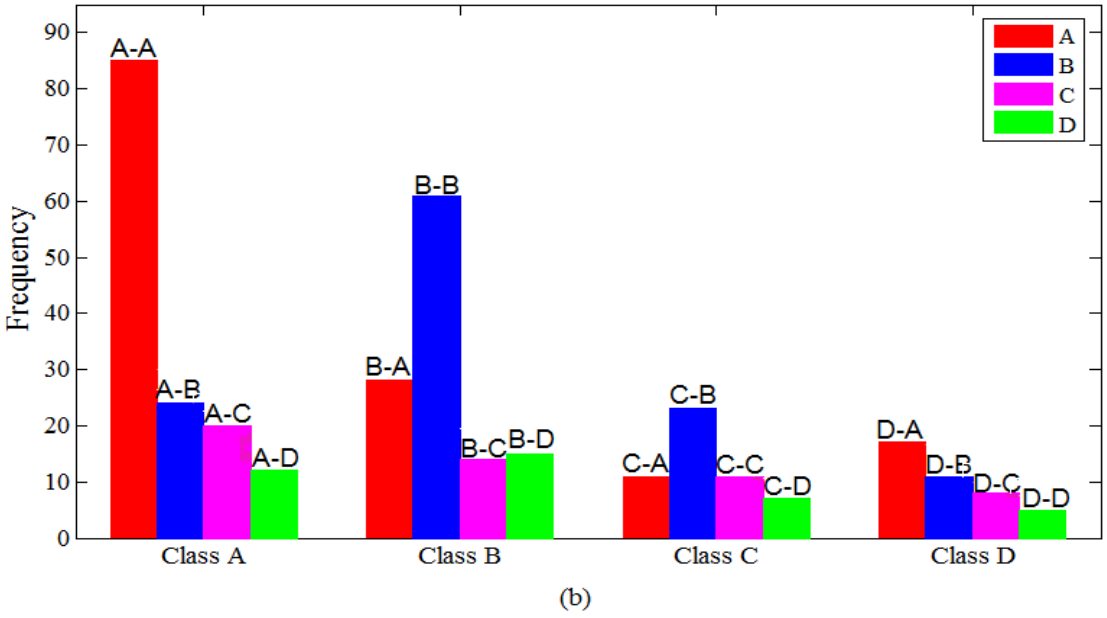


Figure 5.12: (a) Frequency of B_n classes occurring individually and in sets of consecutive days ranging from 2 to 7. (b) Frequency of next day occurrence given the current B_n class of day.

5.5 Comparison with local cloud observations

Since cloud cover is closely related to irradiance, the annual total cloud cover (TCC) distribution was compared with the annual B_n class distribution. In Durban, total cloud cover was recorded by a cloud observer twice a day i.e. at 0600 UTC and 1200 UTC. The TCC in oktas for all cloud levels i.e. low, middle and high at 0600 UTC and 1200 UTC are presented in Figure 5.13 (a) and (b), respectively. Morning TCC (i.e. UTC 0600) between the range of 3 to 5 oktas comprise less than 30 days across all months. However, for afternoon TCC (i.e. UTC 1200) observations this is only the case for 5 oktas. A small number of days across all months have a total cloud coverage of 4 and 5 oktas for the morning and afternoon, respectively. This suggests that morning and afternoon sky coverage, where half the sky or slightly less than half the sky is obscured by clouds, is fairly uncommon for Durban. The months that had the most number of days with a TCC of zero oktas were June and July, which indicates that these months comprised mostly sunny days. This is consistent with the classification and characterization results for Durban, where clustering of the B_n irradiance profiles revealed that the months of June and July had the highest proportion of Class A days. In addition, October, November and December have days with high amounts of TCC and which confirms that Class B (cloudy days) does indeed have a high prevalence during these months. Given that the distribution of cloud cover observations show high correlation with the distribution of irradiance classes, this confirms that the *k*-means clustering produces groups that represent the dominant irradiance patterns in Durban relatively well. Although cloud cover observations are only recorded twice a day, and not as often as irradiance measurements, the information contained in those limited observations nevertheless provides a way of validating the 4 dominant irradiance classes that characterize the irradiance patterns in Durban.

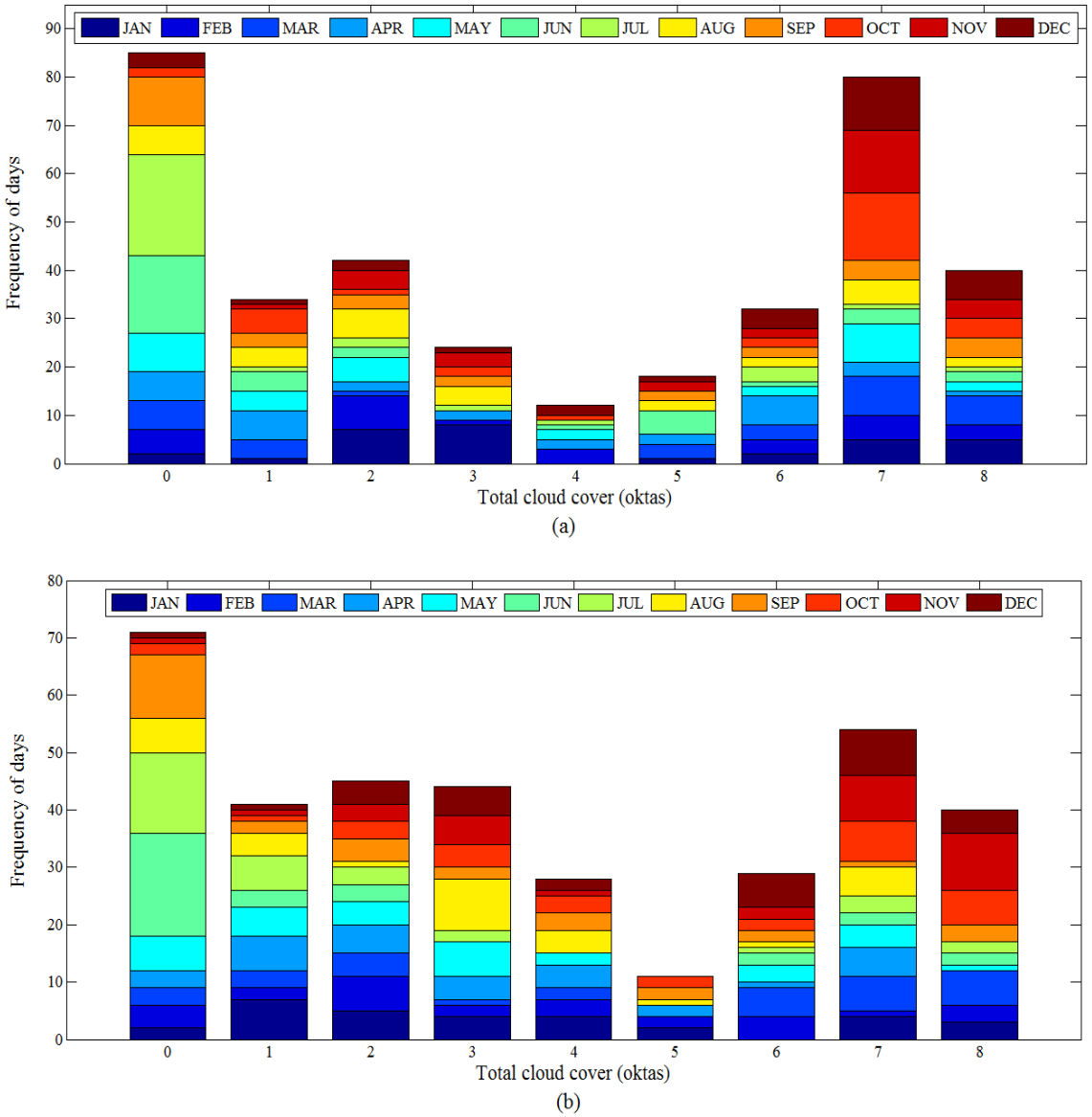


Figure 5.13: Annual distribution of total cloud cover for Durban at (a) 0600 UTC and (b) 1200 UTC.

5.6 Hourly-resolution normalized beam irradiance, $\bar{B}_n$

For each hour, the minute-resolution data was averaged according to equation 2.19. This resulted in a reduced set of profiles of 8 dimensions, where each dimension is the $\bar{B}_n$ value for the hour. The purpose of clustering the hourly-resolution $\bar{B}_n$ was to match the hourly-resolution cloud cover profiles output from the NWP, which was used for forecasting. Secondly, $\bar{B}_n$ was clustered to investigate whether the same clustering patterns from B_n can be re-gained.

Since averaging the minute-resolution profiles to hourly-resolution is in itself a form of data reduction, no PCA is applied to $\bar{B}_n$. Clustering was applied directly to all 8 dimensions. A cluster map produced is a projection onto a 2-d plot that shows a particular view of the data for example, morning and afternoon averages or day averages. For convenience, "*k*-means cluster map" or "cluster map" are used interchangeably when describing the clustered profiles, irrespective of the number of dimensions used for clustering.

The *k*-means cluster map for $\bar{B}_n$ is given in Figure 5.14, which shows morning (8:30-12:30) average of $\bar{B}_n$ on the horizontal axis and afternoon (12:30-16:30) average on the vertical axis. Classes A and B are compact and widely separated whereas Classes C and D, in the region between A and B, are less compact and have members at their edges that are closer on average to A and B than to their own cluster, resulting in negative SI for those members. A minor difference between the $\bar{B}_n$ and B_n clustering is that the number of profiles with negative SI decreased from 19 for B_n to 16 for $\bar{B}_n$. Table 5.2 summarizes the clustering information.

Table 5.2: Summary of $\bar{B}_n$ clustering.

Class	Cluster	Frequency of days	Proportion	$\overline{SI}_C$	$\overline{SI}_C < 0$
A	1	135	37%	0.79	0
B	2	124	34%	0.79	0
C	3	65	18%	0.33	9
D	4	41	11%	0.25	7

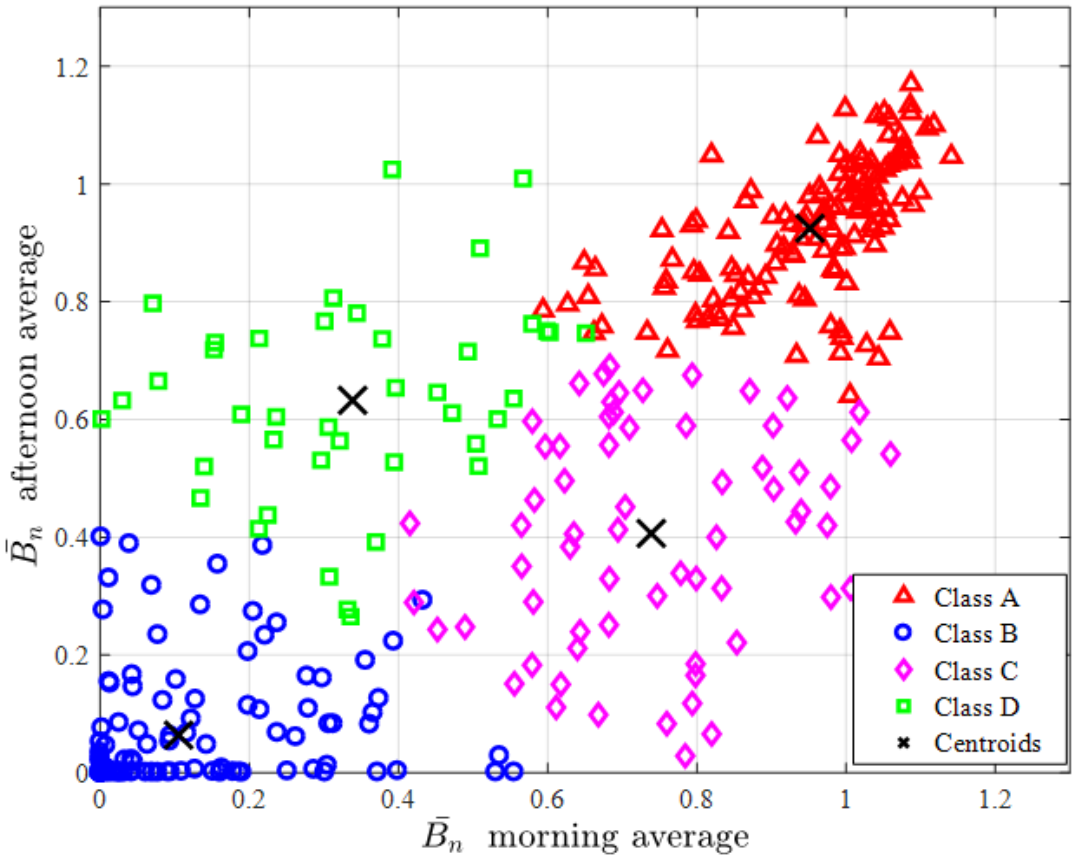


Figure 5.14: $\bar{B}_n$ cluster map of the morning average (8:30-12:30) and afternoon average (12:30-16:30). Class A days have high morning and afternoon averages, whereas Class B is low for both. Class C has high morning averages and low afternoon averages. Class D is the opposite with low morning averages and high afternoon averages.

Class mean profiles for $\bar{B}_n$, denoted as $\langle \bar{B}_n \rangle$, are shown in Figure 5.15. The mean profiles are similar to those the profiles of B_n. Class A contains days that have high levels of $\bar{B}_n$ throughout the day. Alternatively, Class B has low levels of $\bar{B}_n$. For Class C $\bar{B}_n$ is high until midday and thereafter is low. Class D is the opposite of Class C. In addition to $\bar{B}_n$, there are a set of associated hourly-resolution profiles for the normalized diffuse irradiance $\bar{D}_n$. These are given in Figure 5.16.

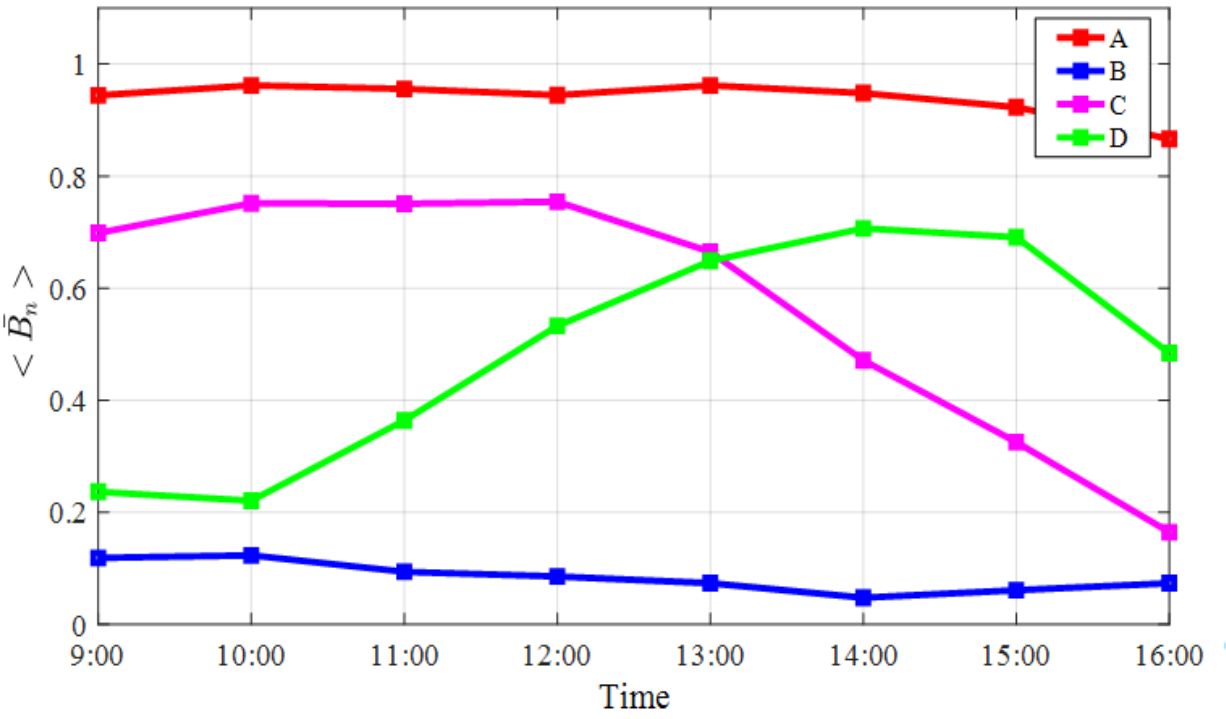


Figure 5.15: Mean profiles of the $\bar{B}_n$ classes. Class A has high levels of $\bar{B}_n$ all day, while Class B is the opposite. $\bar{B}_n$ is high in the morning and low in the afternoon for Class C, and vice-versa for Class D. All profiles are similar to profiles of B_n.

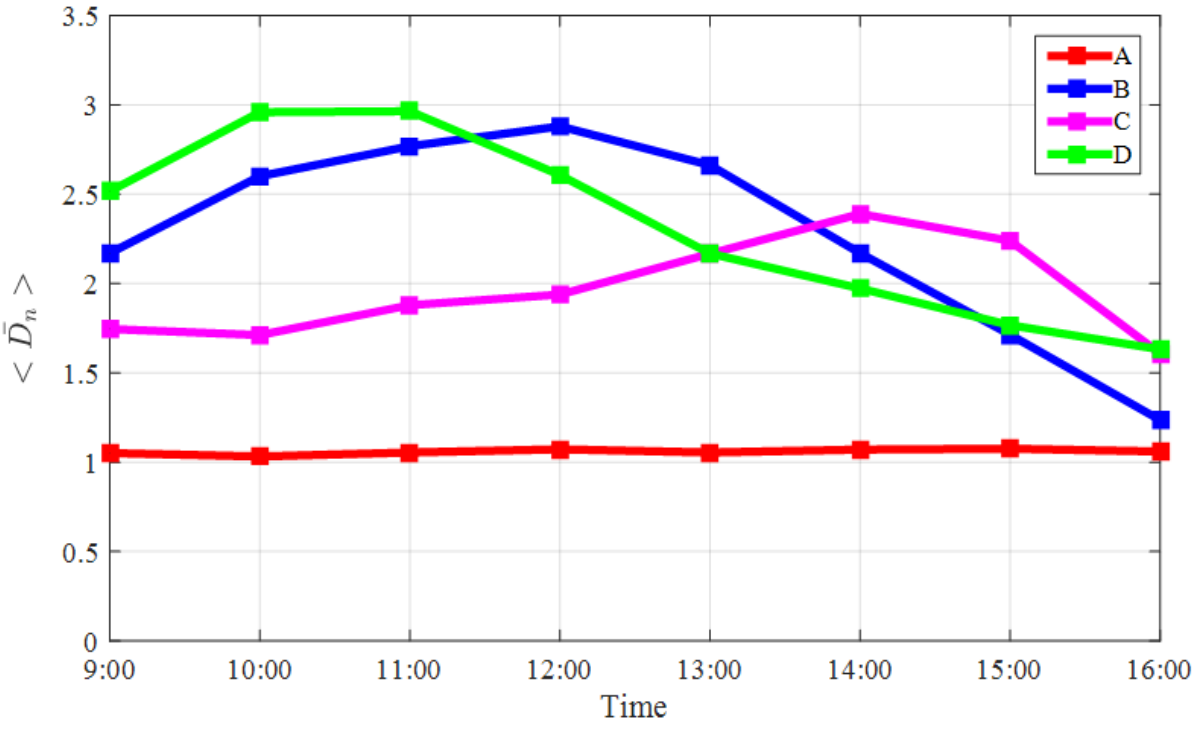


Figure 5.16: Mean profiles of the $\bar{D}_n$ classes. These profiles are similar to those of minute-resolution D_n.

For forecasting it is important to have knowledge of the range of uncertainty. The hourly-resolution $\bar{B}_n$ class mean profiles in the form of box plots are shown in Figures 5.17 and 5.18, calculated for each hour. The more compact clusters (A and B) have smaller interquartile (i.e. low variability) range than the less compact clusters (C and D), which have a larger interquartile range (i.e. high variability). This is because sunny and cloudy conditions have low variability in the B_n as compared to days with mixed conditions (Classes C and D). A forecast that places a day into a class will thus be predicted to have an irradiance profile similar to that class mean profile within an uncertainty corresponding to the class standard deviation. It is therefore expected that forecasts of A and B class will have less uncertainty than for C and D.

The associated $\langle \bar{D}_n \rangle$ profiles with their corresponding box plot for each $\bar{B}_n$ class are presented in Figures 5.19 and 5.20. The $\bar{D}_n$ profiles have diurnal trends that are the inverse of $\bar{B}_n$, where Class A $\bar{D}_n$ is very low throughout the day at 1. Class B has high $\bar{D}_n$ due to the scattering of the beam fraction by clouds. Classes C and D have high $\bar{D}_n$ during the cloudy periods, however the $\bar{D}_n$ levels are not as high as in the case of Class B thus indicating less scattering. Class A has a rather small range of uncertainty as compared to Classes B, C and D.

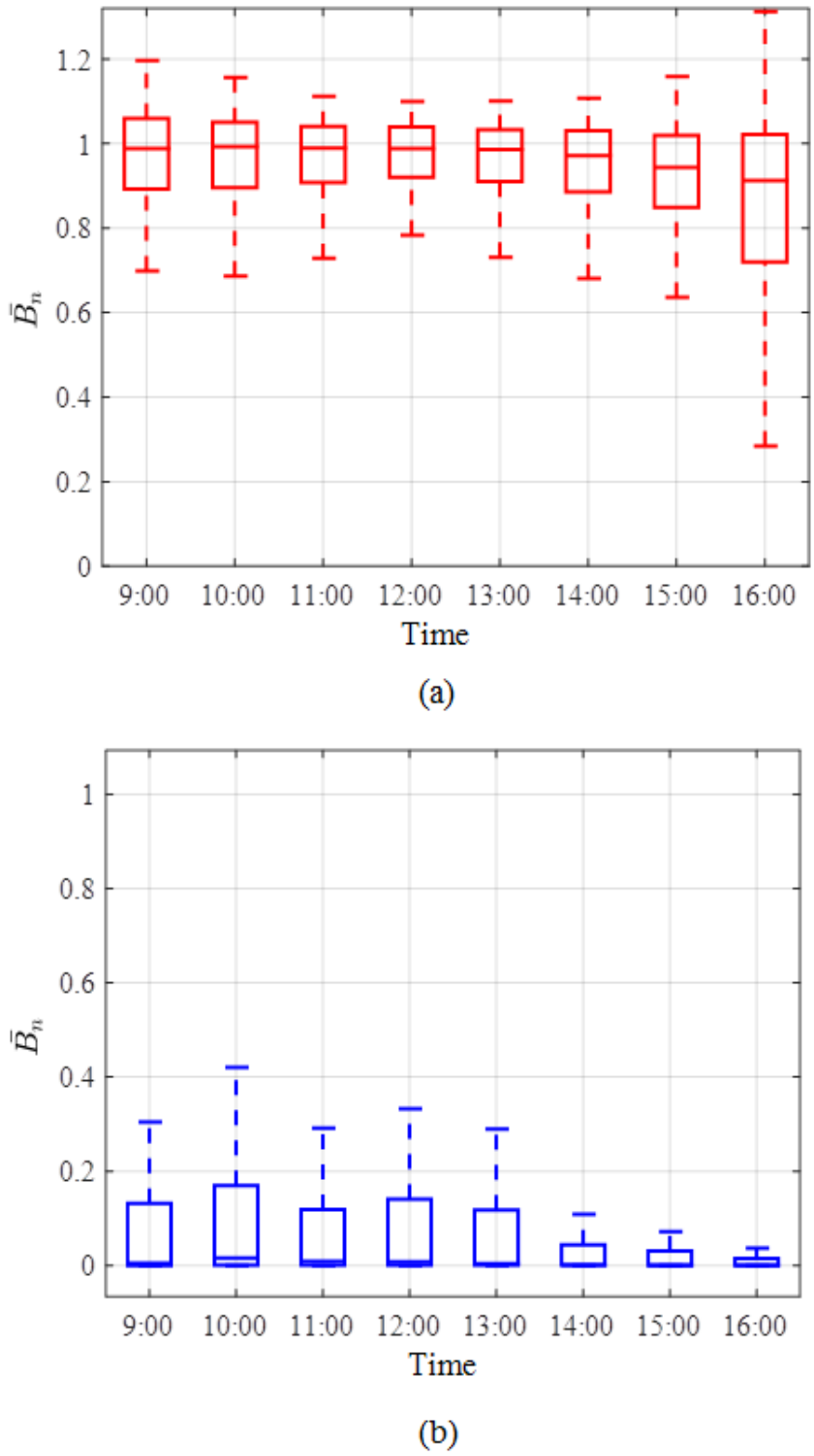
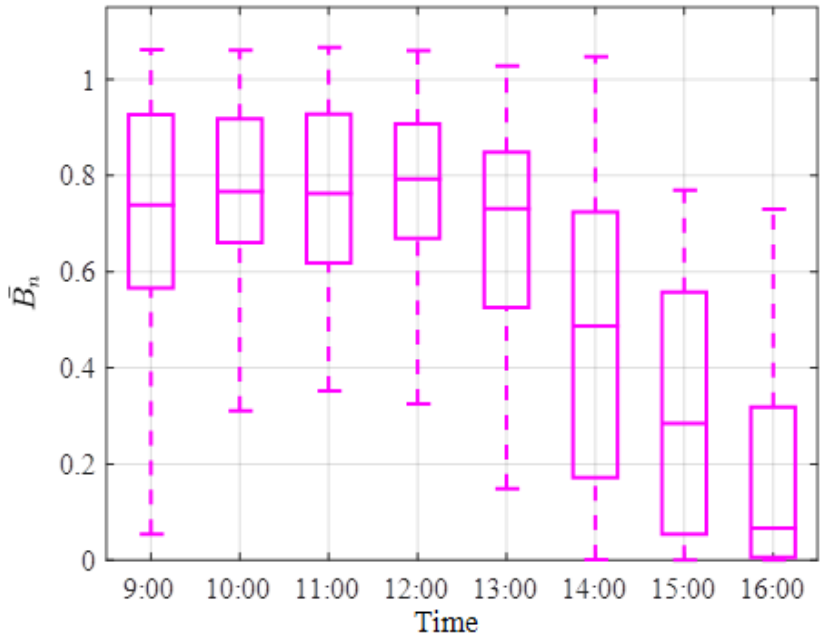
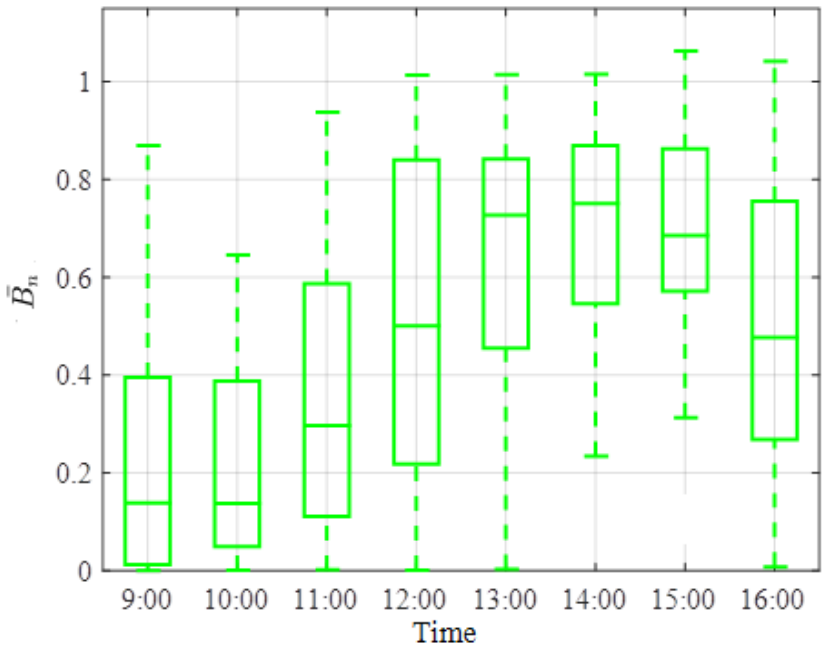


Figure 5.17: Box plots for hourly class mean profiles of $\bar{B}_n$ for (a) sunny (Class A), (b) cloudy days (Class B). Classes A and B have relatively small interquartile range in comparison to Classes C and D.



(a)



(b)

Figure 5.18: Box plots for hourly class mean profiles of $\bar{B}_n$ for (a) sunny AM-cloudy PM (Class C) and (d) cloudy AM-sunny PM days (Class D). Classes C and D have relatively large interquartile range as compared to Classes A and B.

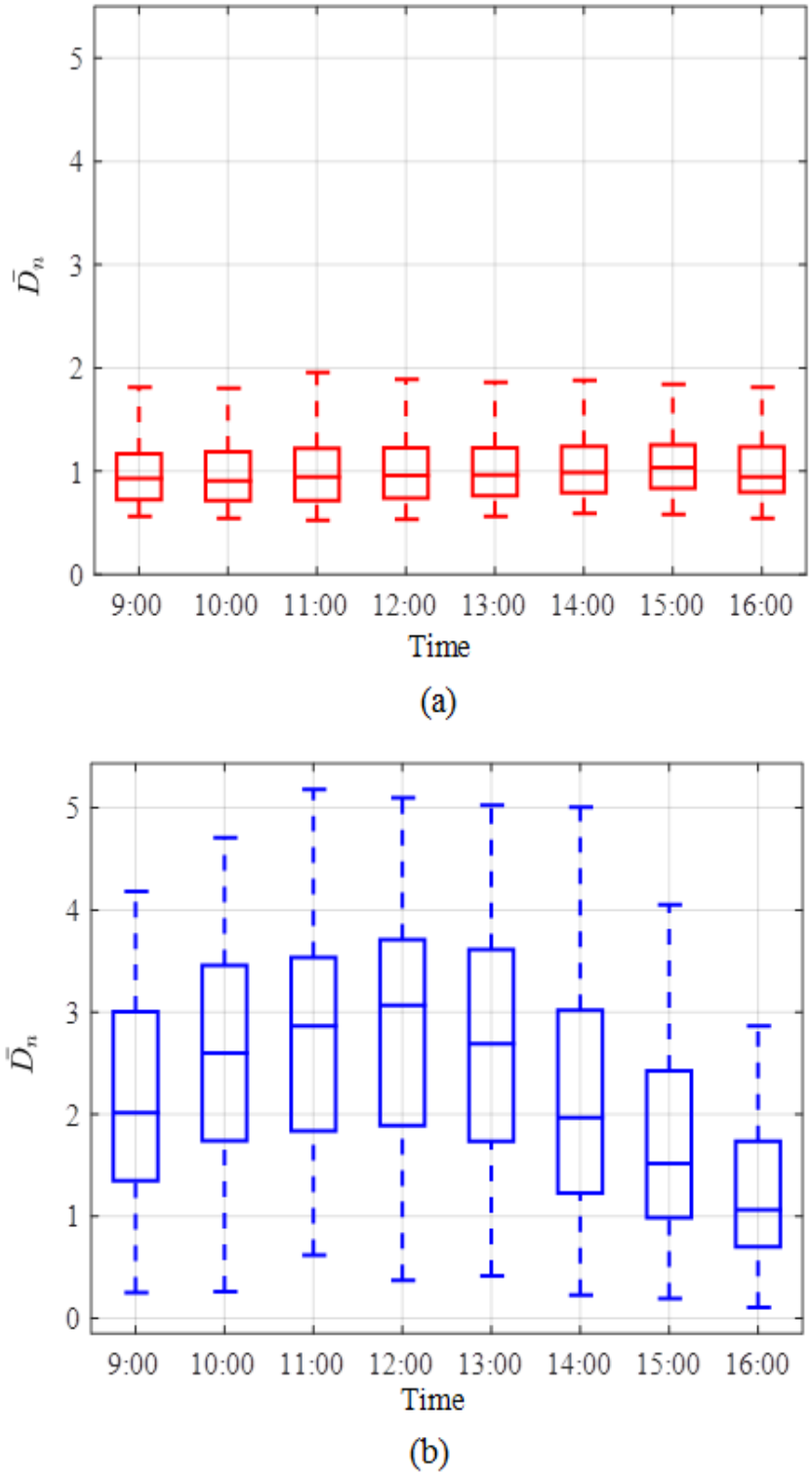


Figure 5.19: Box plots for hourly class mean profiles of $\bar{D}_n$ for (a) sunny (Class A), (b) cloudy days (Class B). Classes A and B have relatively small interquartile range as compared to Classes C and D.

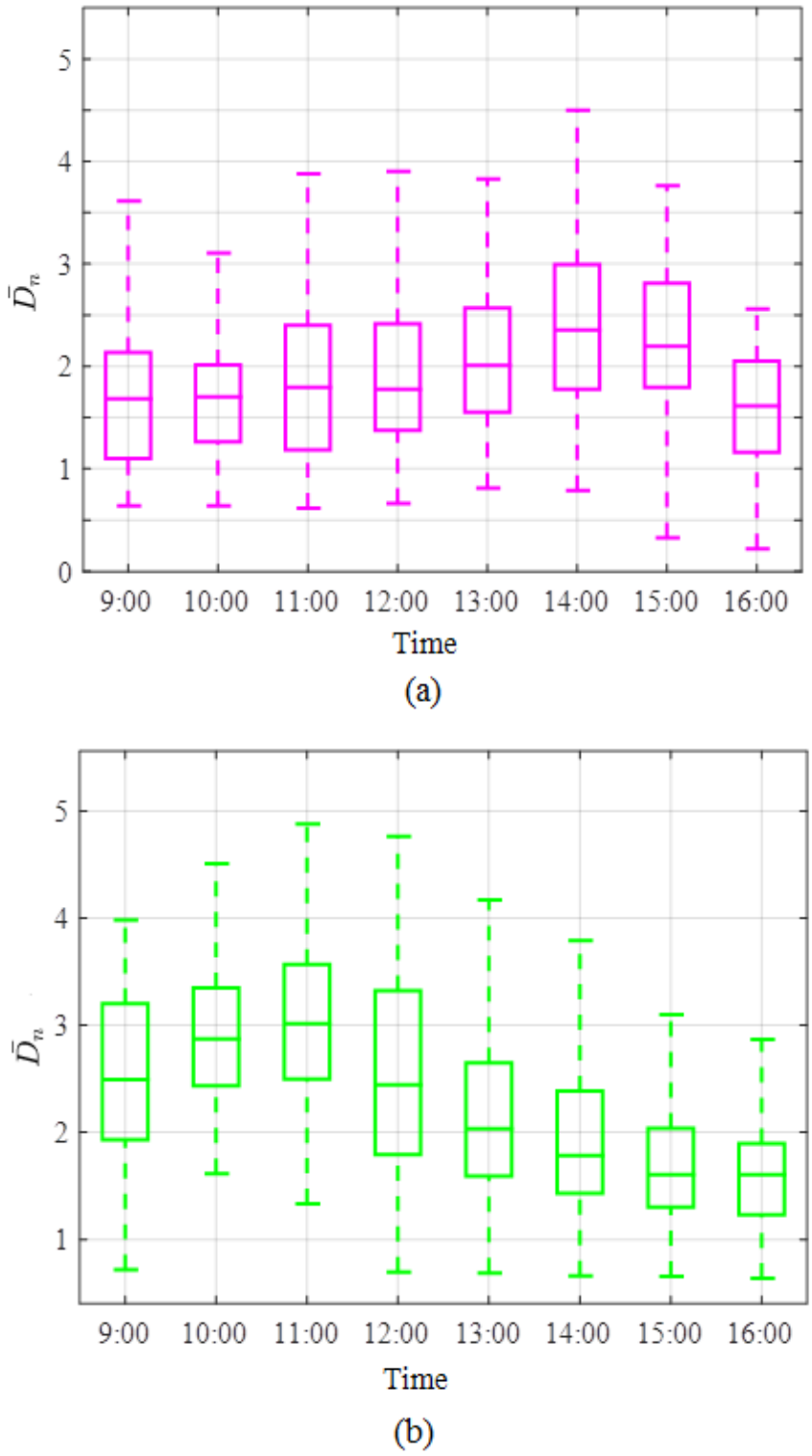


Figure 5.20: Box plots for hourly class mean profiles of $\bar{D}_n$ for (a) sunny AM-cloudy PM (Class C) and (d) cloudy AM-sunny PM days (Class D). Classes C and D have relatively large interquartile range as compared to Classes A and B.

5.7 Hourly-resolution variability, V_B

Variability in B_n (V_B) was investigated since V_B may contain information about B_n that may have been lost through averaging. This led to the clustering of V_B, since clustering could find some pattern that may exist in V_B that is potentially useful for forecasting.

Variability of irradiance on the diurnal scale i.e. movement of the Sun during the day is deterministic and can therefore be easily predicted (Kleissl, 2013). Alternatively, variability due to atmospheric effects, in particular clouds, is difficult to predict especially on the minute-resolution timescale. Due to this rapid fluctuation in the minute-resolution data being difficult to predict, an average value for the variability that takes into account these fluctuations was used. For this thesis, the variability in B_n defined in Chapter 2 was used as a measure for the variability over the hour.

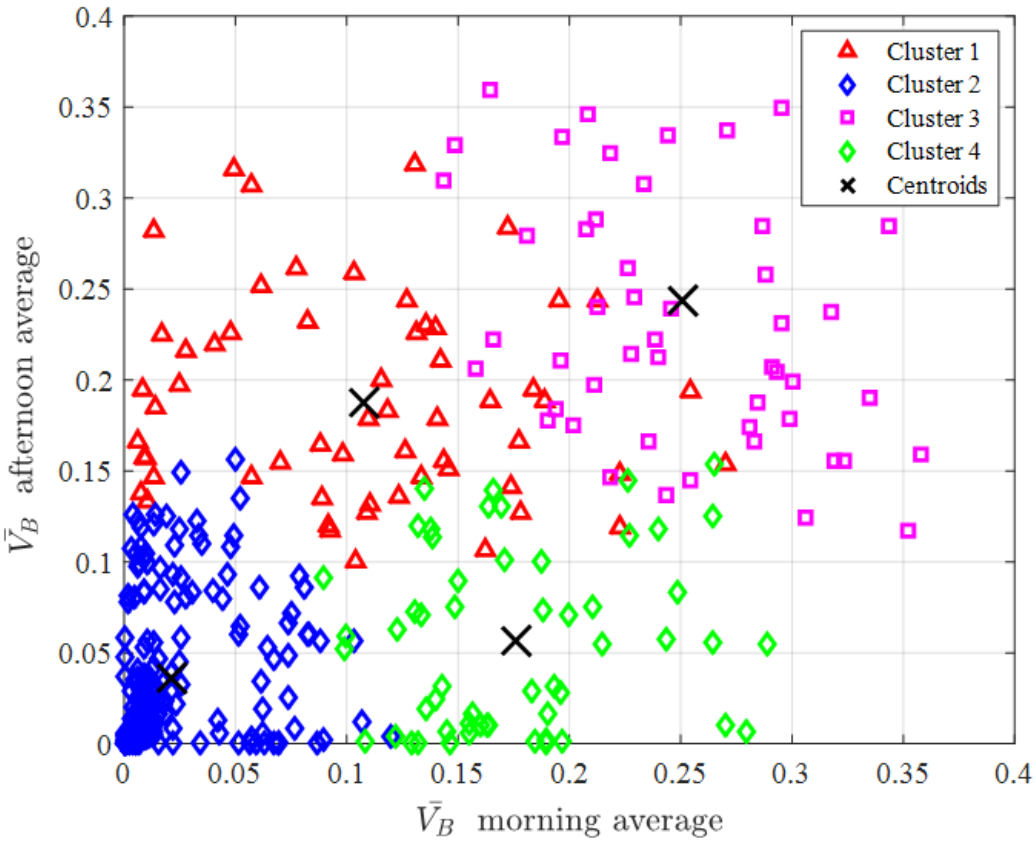


Figure 5.21: $\bar{V}_B$ cluster map of the morning average and afternoon average. Cluster 1 contains days that have low morning and afternoon average variability. Days in Cluster 2 have low morning and high afternoon variability. Cluster 3 and 4 have a combination of low and high variability in the mornings and afternoons.

Similar to B_n, there were 8 V_B dimensions and no PCA was applied. The cluster map for the morning and afternoon average of V_B is shown in Figure 5.21. Similar to $\bar{B}_n$, there are two clusters with low and high values for V_B, and two clusters with intermediate V_B values. Table 5.3 summarizes the V_B clustering. The mean profiles of the clusters in Figure 5.21 are shown in Figure 5.22. From the mean $\langle \bar{V}_B \rangle$ profiles it is not clear which $\bar{B}_n$ class of days A, B, C or D are contained in the V_B clusters. The profile with low $\langle \bar{V}_B \rangle$ throughout the day can be associated with Class A and B days i.e. sunny or cloudy all day, respectively, since the sky conditions are persistent throughout the day and hence $\langle \bar{V}_B \rangle$ will be low. The remaining V_B profiles do not show a pattern that can be correlated with a $\bar{B}_n$ class. For this reason, the associated mean $\bar{B}_n$ profiles are given in Figure 5.23. This set of $\bar{B}_n$ profiles lack the distinct diurnal patterns as compared to the profiles in Figure 5.15, and all four of these profiles lie in the range 0.27-0.65. This is because days that were separated originally in a $\bar{B}_n$ Class A, B, C or D are now mixed, and based on this particular clustering solution, one $\bar{B}_n$ class may contain a combination of Class A, B, C or D days. This results in the average over the $\langle \bar{B}_n \rangle$ class members having no distinct diurnal pattern.

Table 5.3: Summary of V_B clustering.

Cluster	**Frequency of days**	**Proportion**	$\overline{\boldsymbol{SI}_{\boldsymbol{C}}}$	$\overline{\boldsymbol{SI}_{\boldsymbol{C}}} < 0$
1	59	16%	0.08	20
2	198	54%	0.74	0
3	53	15%	0.17	11
4	55	15%	0.20	10

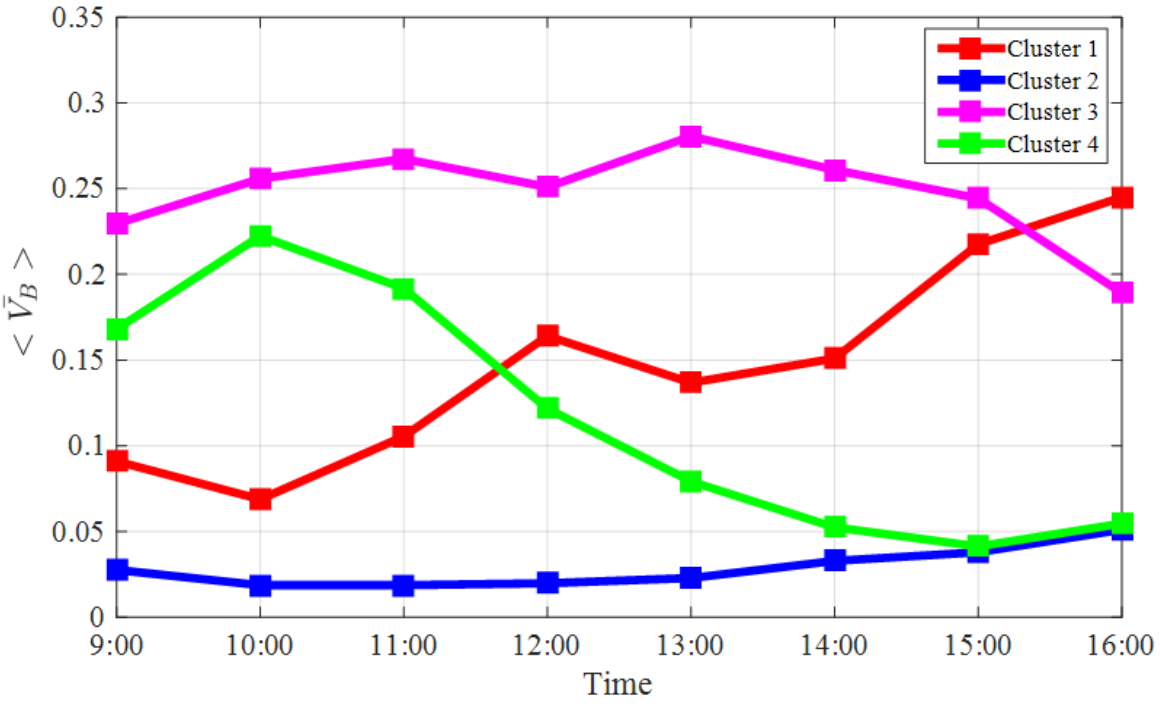


Figure 5.22: Mean profiles for V_B clusters.

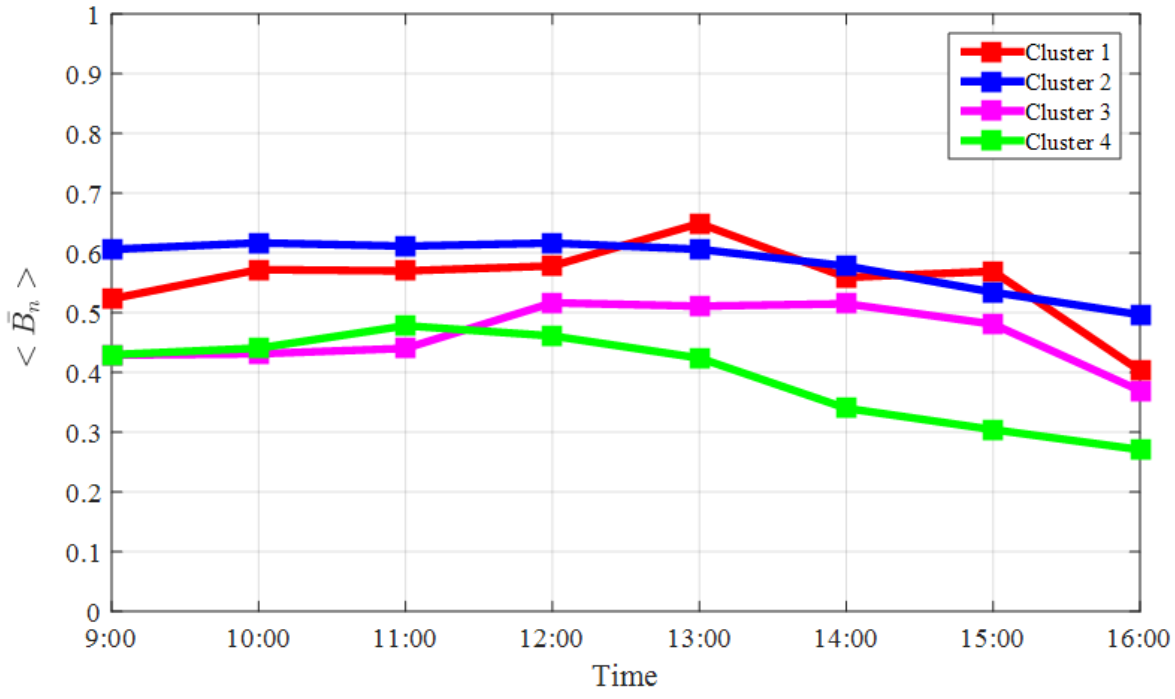


Figure 5.23: Associated mean profiles for $\bar{B}_n$ for V_B clusters.

5.8 Combination of $\{\bar{B}_n, V_B\}$

As mentioned earlier, clustering of $\bar{B}_n$ combined with V_B denoted as $\{\bar{B}_n, V_B\}$ was investigated. Clustering of this combination was to investigate whether $\bar{B}_n$ with its variability has any effect on the clustering structure and more specifically, if stronger clustering patterns emerge. The combination of the hourly values of $\{\bar{B}_n, V_B\}$ resulted in 16 dimensions and *k*-means clustering was applied directly to the 16-dimensional set. Table 5.4 summarizes the clustering information for the $\{\bar{B}_n, V_B\}$ combination. The cluster maps for the $\{\bar{B}_n, V_B\}$ combinations for the morning and afternoon averages are shown in Figures 5.24 and 5.25. The day averages of $\bar{B}_n$ and V_B are shown in Figure 5.26.

Table 5.4: Summary of $\{\bar{B}_n, V_B\}$ clustering.

Cluster	Frequency of days	Proportion	$\overline{SI}_C$	$\overline{SI}_C < 0$
1	54	15%	0.11	20
2	59	16%	0.33	5
3	131	36%	0.78	0
4	121	33%	0.78	0

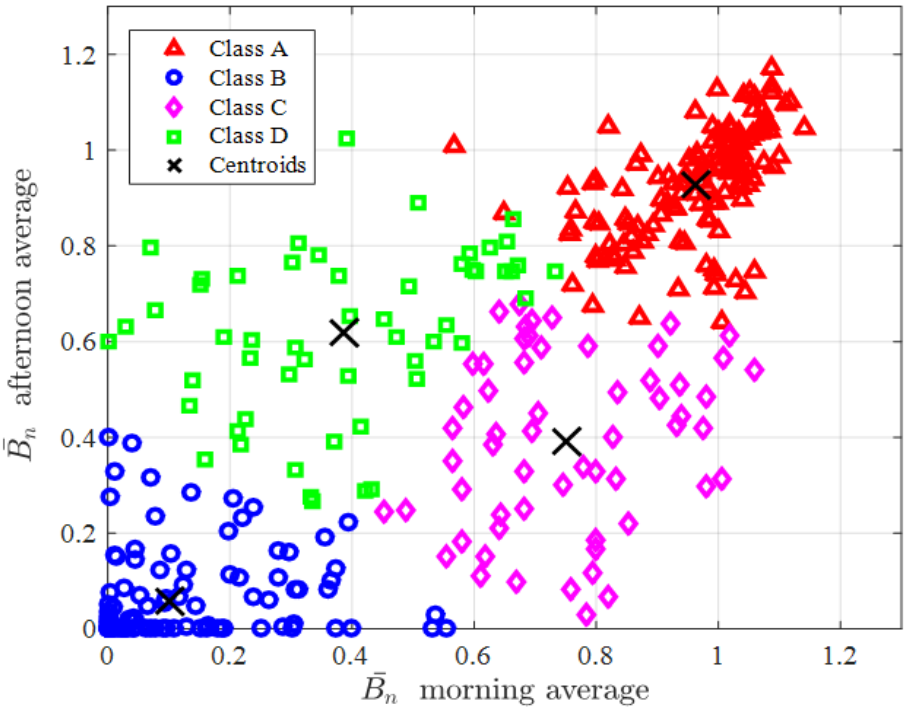


Figure 5.24: Cluster map of the morning and afternoon $\bar{B}_n$ average for the combination of $\{\bar{B}_n, V_B\}$.

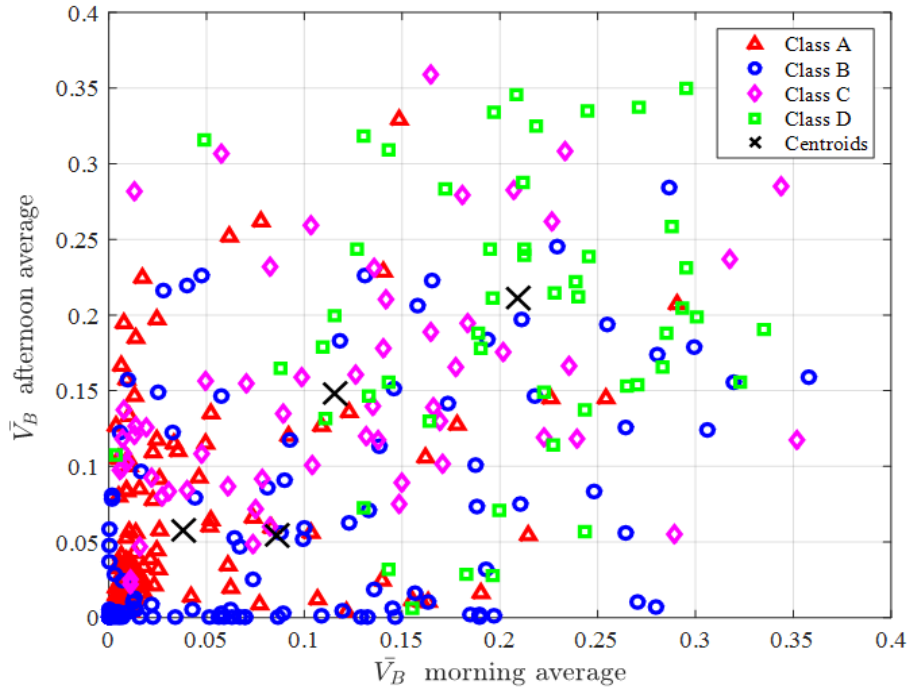


Figure 5.25: Cluster map of the morning and afternoon V_B average for the combination of $\{\bar{B}_n, V_B\}$.

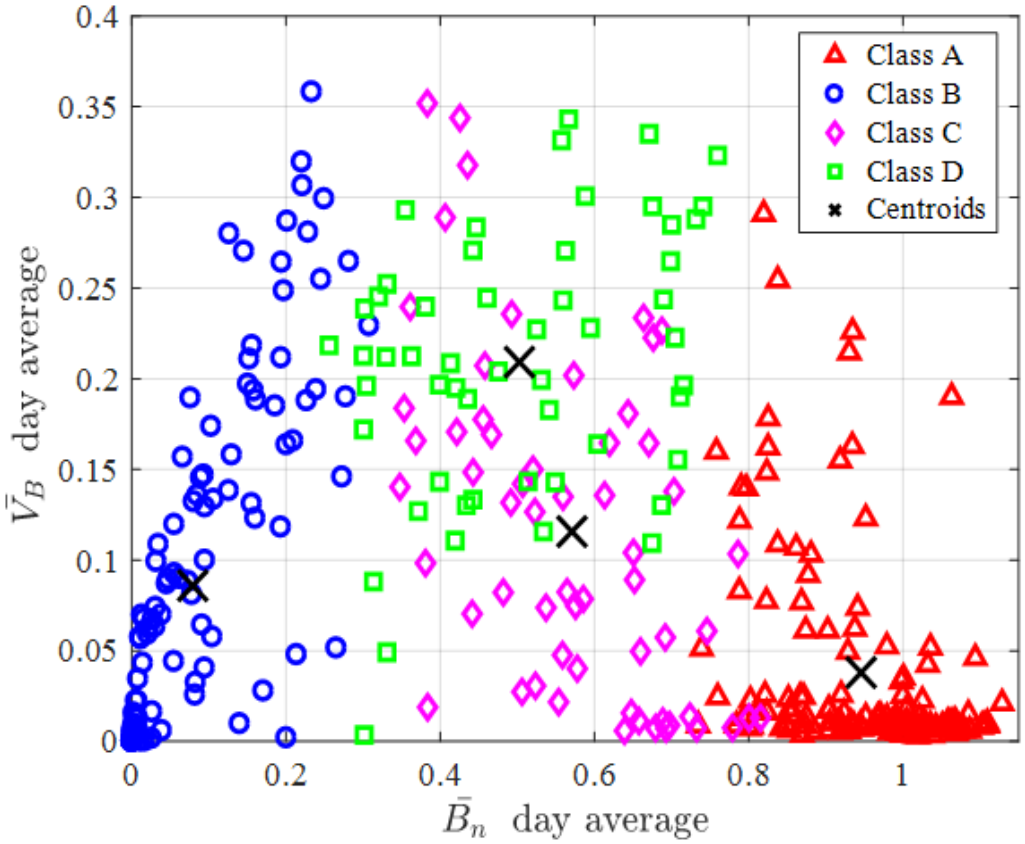


Figure 5.26: Cluster map of $\bar{B}_n$ and V_B averaged over all hours of the day.

The mean profiles for $\bar{B}_n$ and V_B are shown in Figure 5.27 and 5.28. The $\bar{B}_n$ profiles are similar to those produced by clustering $\bar{B}_n$ on its own (Figure 5.15). The mean profiles for V_B show that Classes A and B have fairly similar trends i.e. low V_B throughout the day. An exception is the last hour of Class A, where V_B tends to show a slight increase. Classes C and D have higher V_B levels than A and B, which is expected due to the change in sky conditions for these types of days.

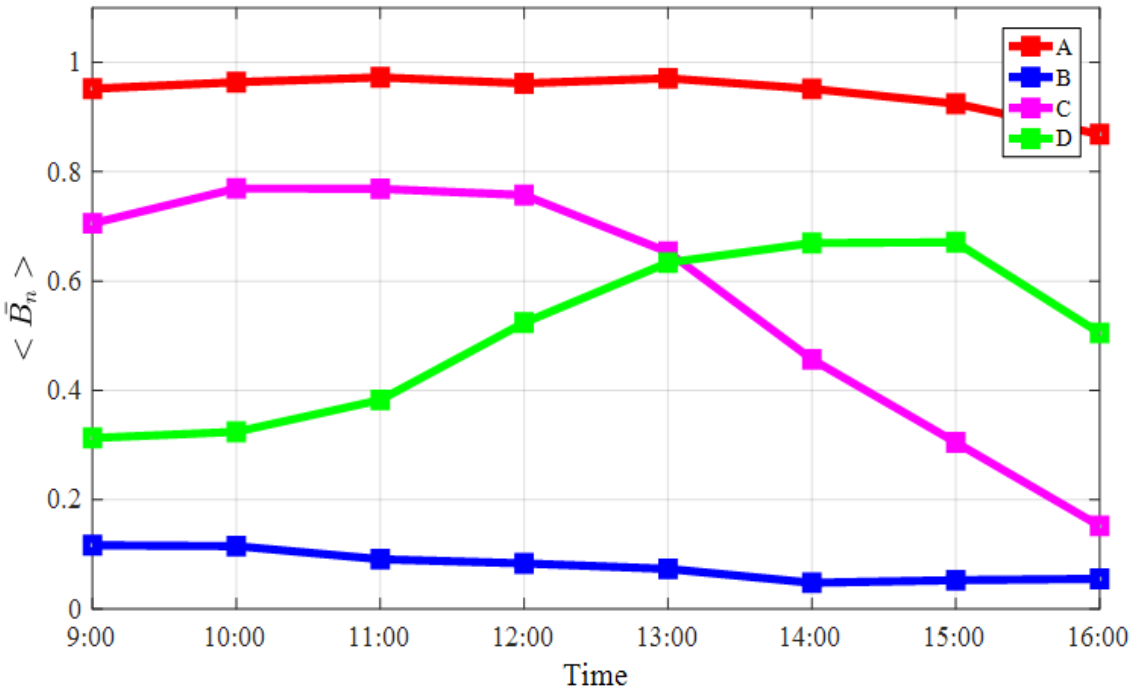


Figure 5.27: Mean profiles for $\bar{B}_n$ when combined with V_B.

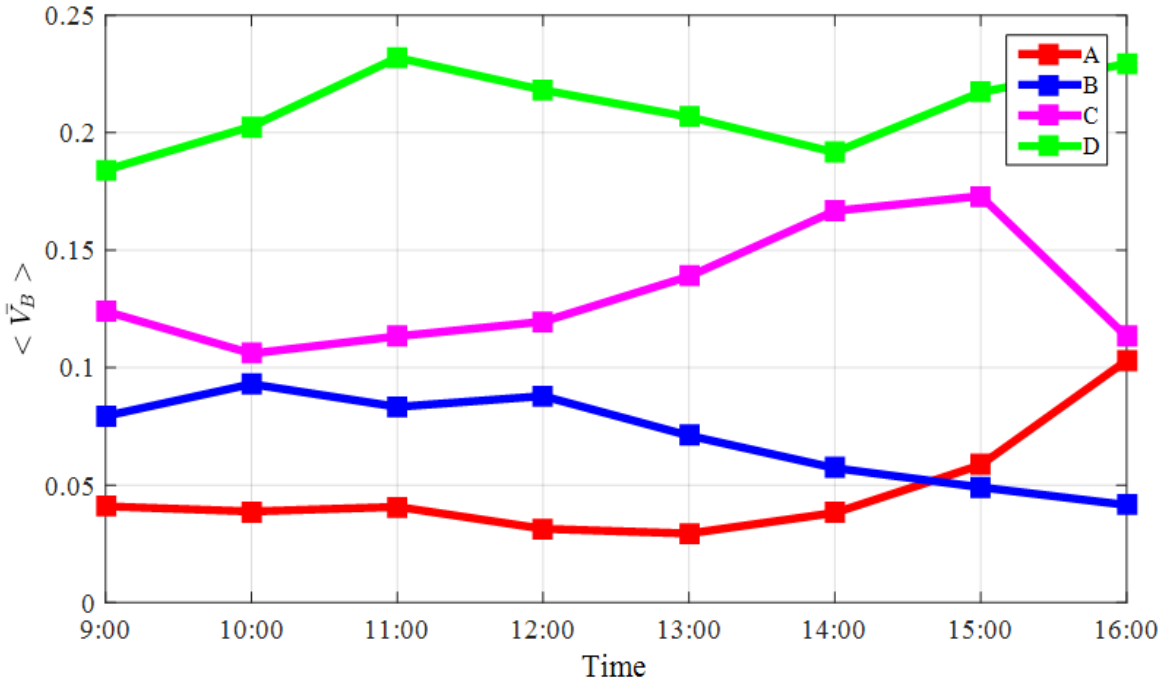


Figure 5.28: Mean profiles for V_B when combined with $\bar{B}_n$.

5.9 Combination of $\{\bar{B}_n, \bar{D}_n\}$

To investigate whether the additional information of diffuse irradiance yields a different clustering solution, $\bar{B}_n$ and $\bar{D}_n$ were combined, denoted as $\{\bar{B}_n, \bar{D}_n\}$, and clustered also as a 16-dimensional set. The cluster map of the morning and afternoon averages of the $\{\bar{B}_n, \bar{D}_n\}$ combination are presented in Figures 5.29 and 5.30, respectively. When combined with $\bar{D}_n$, the cluster map for $\bar{B}_n$ results in Class A being relatively compact and well separated, but Classes B, C and D are weakly compact. Alternatively, the cluster map for $\bar{D}_n$ shows an improvement in compactness in all classes.

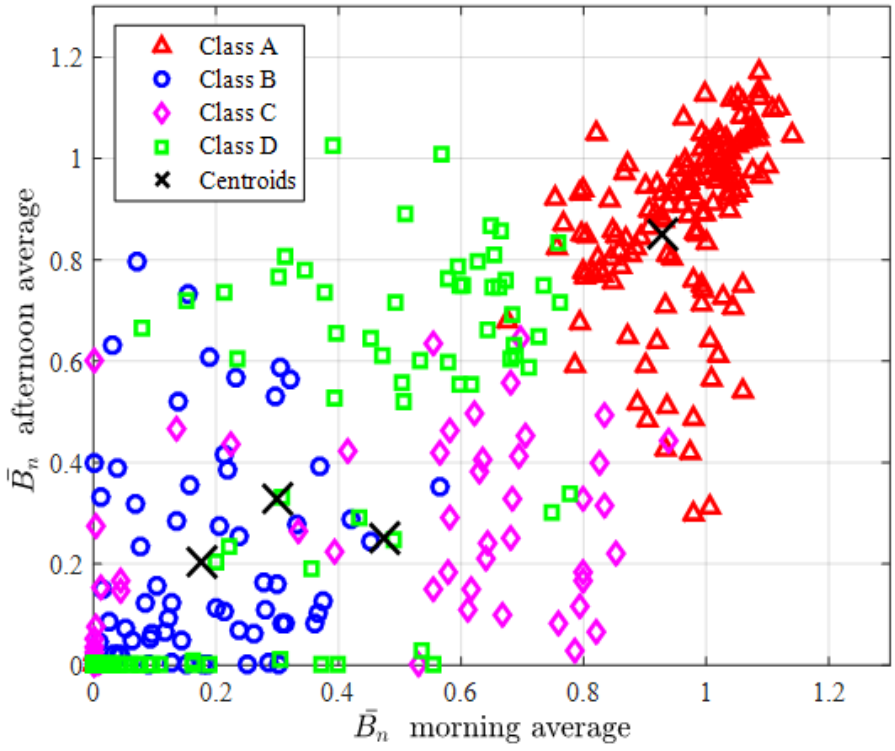


Figure 5.29: Cluster map of $\bar{B}_n$ when combined with $\bar{D}_n$. Class A is relatively compact and well separated, but Classes B, C and D are weakly compact.

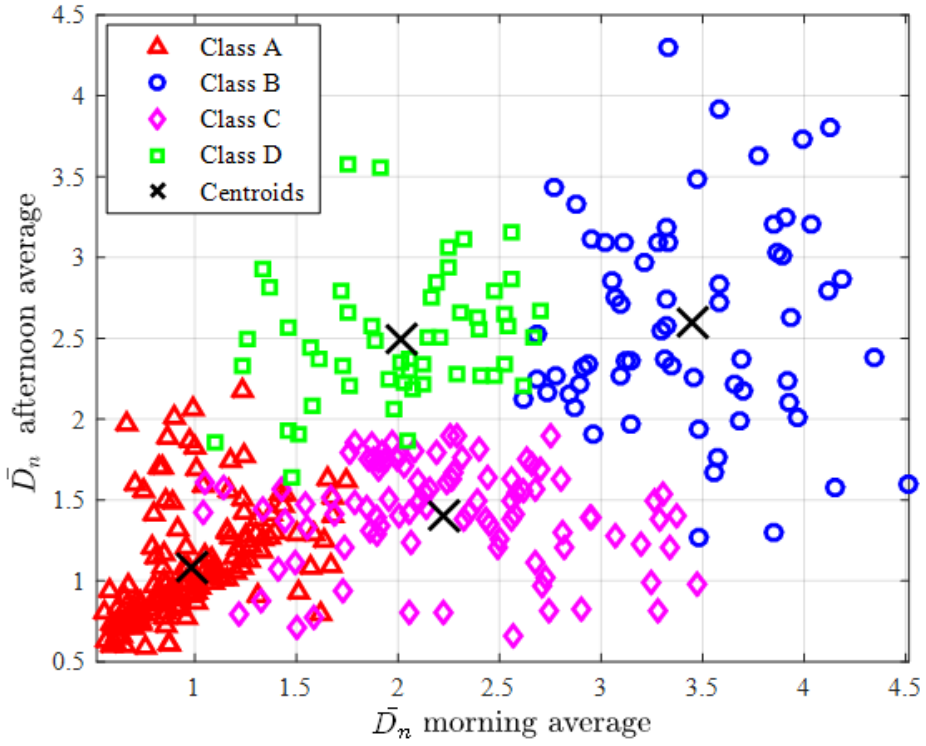


Figure 5.30: Cluster map of $\bar{D}_n$ when combined with $\bar{B}_n$.

Figures 5.31 and 5.32 show the class mean profiles for $\bar{B}_n$ and $\bar{D}_n$, respectively. It is quite evident that the $\bar{B}_n$ classes have a different set of members and therefore produce different mean profiles as compared to Figure 5.15. Classes A and B are regained even when combined with $\bar{D}_n$ however, Classes C and D are weak in their $\bar{B}_n$ profiles and tend to lose the distinct shapes that were present in Figure 5.15. On the other hand, Classes C and D are stronger in their $\bar{D}_n$ profiles with their afternoons being strongly separated. The mean profiles for $\bar{D}_n$ are fairly consistent with those in Figure 5.16, in the sense that they represent the same diurnal patterns in the diffuse irradiance.

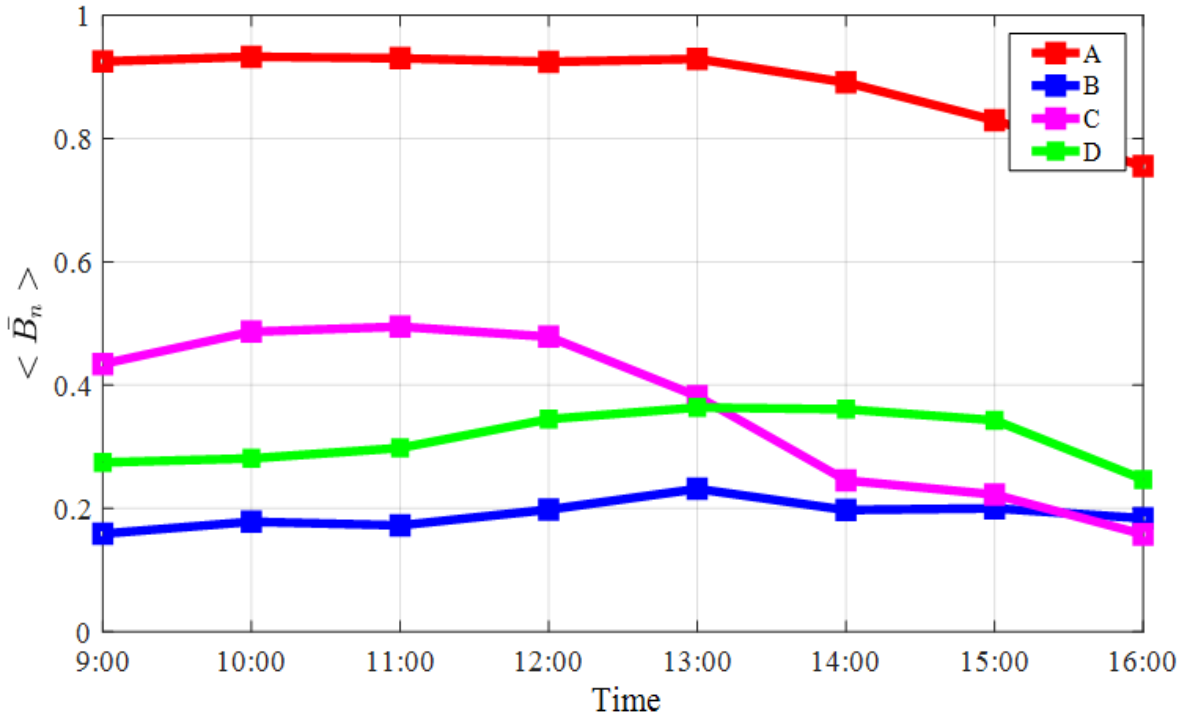


Figure 5.31: Mean profiles for $\bar{B}_n$ when combined with $\bar{D}_n$.

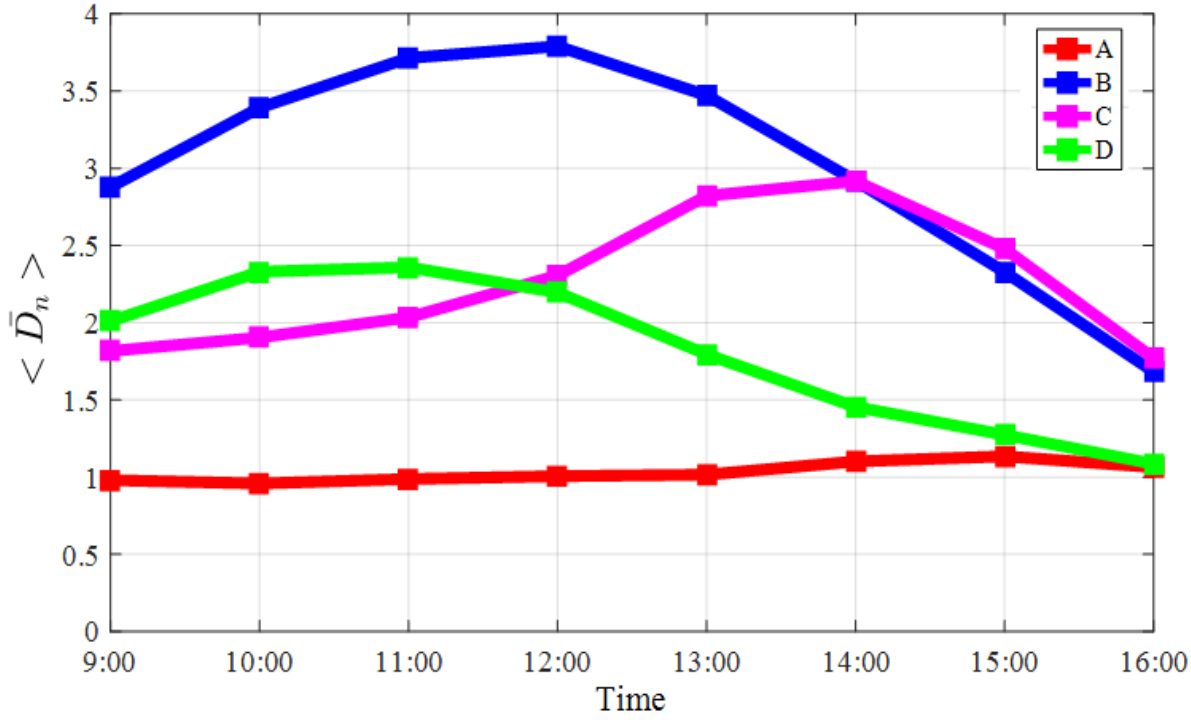


Figure 5.32: Mean profiles for $\bar{D}_n$ when combined with $\bar{B}_n$.

Chapter 6

Forecasting using classes

This chapter describes forecasting of irradiance profiles using classes. Day forecasts are presented using two methods, both of which depend on cloud cover output from the NWP. Intra-day forecasting, in particular, hourly forecasts for $\bar{B}_n$ and $\bar{D}_n$ are presented. All forecasting methods depend on the clustering results and the classes of $\bar{B}_n$ and $\bar{D}_n$ that were established through clustering in Chapter 5. Testing of forecasting was done using a set of 100 days during January to June in 2017. The test set included radiometric and cloud cover profiles. Forecasting results are presented with the relevant success rates and error metrics.

6.1 Day forecasts

The class mean profiles of $\bar{B}_n$, presented in Chapter 5, were used for forecasting the class of day. More specifically, forecasting was done using a day-ahead forecast of Q to forecast the class of the day, and then using $\langle \bar{B}_n \rangle$ for that class as the forecast of the $\bar{B}_n$ profile. Although Q was specified on the hour in "clock time" (SAST), and hence offset from solar time, the difference of at most 20 minutes was small compared with the one-hour resolution of the Q profile and was not significant. Forecasting was done using two methods. The first uses classes of Q found by k-means clustering. The second method uses a simple set of decision rules referred to as the "50% rule".

6.1.1 Clustering of hourly-resolution cloud cover, Q

Forecasting day-ahead irradiance used cloud cover forecasts from the NWP that were clustered by the k-means method. Hourly cloud cover profiles, Q, were recorded for 243 days in the year 2016.

As mentioned earlier, these were obtained from AccuWeather, a public weather-service provider, to produce a daily profile from 9:00 to 16:00 SAST (South African Standard Time). The Q output from AccuWeather makes use of the GFS NWP model run by National Oceanic and Atmospheric Administration (NOAA). Q profiles are available at hourly temporal resolution for at least 12 hours ahead. Therefore, at the start of the day, the Q profile for the entire day ahead at each hour is available.

Clustering of Q profiles from the NWP output was a novel aspect of this thesis, and was done so that the classes of Q could be used to forecast the irradiance class. The forecast of Q is assigned to one of the classes, which in turn is associated with one of the classes of $\bar{B}_n$. However, in order to use classes of Q to forecast the irradiance class, it was necessary to have four classes of cloud cover. Because low $\bar{B}_n$ is correlated with high Q, it was expected that they should exhibit similar clustering considering that NWP is designed to model the weather pattern as accurately as possible. Nevertheless, k was varied from 2 to 10 as for irradiance profiles in order to check whether k = 4 was a good solution. It was found that k = 4 was indeed a good solution, with $\overline{SI}_{TOT}$ = 0.74. Since Q profiles were at hourly intervals, similar to $\bar{B}_n$, the number of dimensions were 8, and so no pre-processing was applied. k-means clustering was applied directly to the 243 days of the 8-dimensional profiles.

The cluster map of the morning and afternoon average of Q is given in Figure 6.1, with morning average of Q on the horizontal axis and the afternoon average on the vertical axis. The four clusters have cloud cover conditions that are associated with the $\bar{B}_n$ classes and are given the same labels, but with primes. Thus the Q classes are Class A′: low Q all day, Class B′: high Q all day, Class C′: low Q in morning and high Q in afternoon, and Class D′: high Q in morning and low Q in afternoon. Class A′ correspond to sunny days, Class B′ to cloudy days, Class C′ to days that are sunny in the morning and cloudy in the afternoon and Class D′ to days that are cloudy in the morning and sunny in the afternoon. Table 6.1 summarizes the results of clustering of daily Q profiles. The total number of days with $\overline{SI}_C < 0$ were 6. The mean Q profiles of the clusters in Figure 6.1 are given in Figure 6.2.

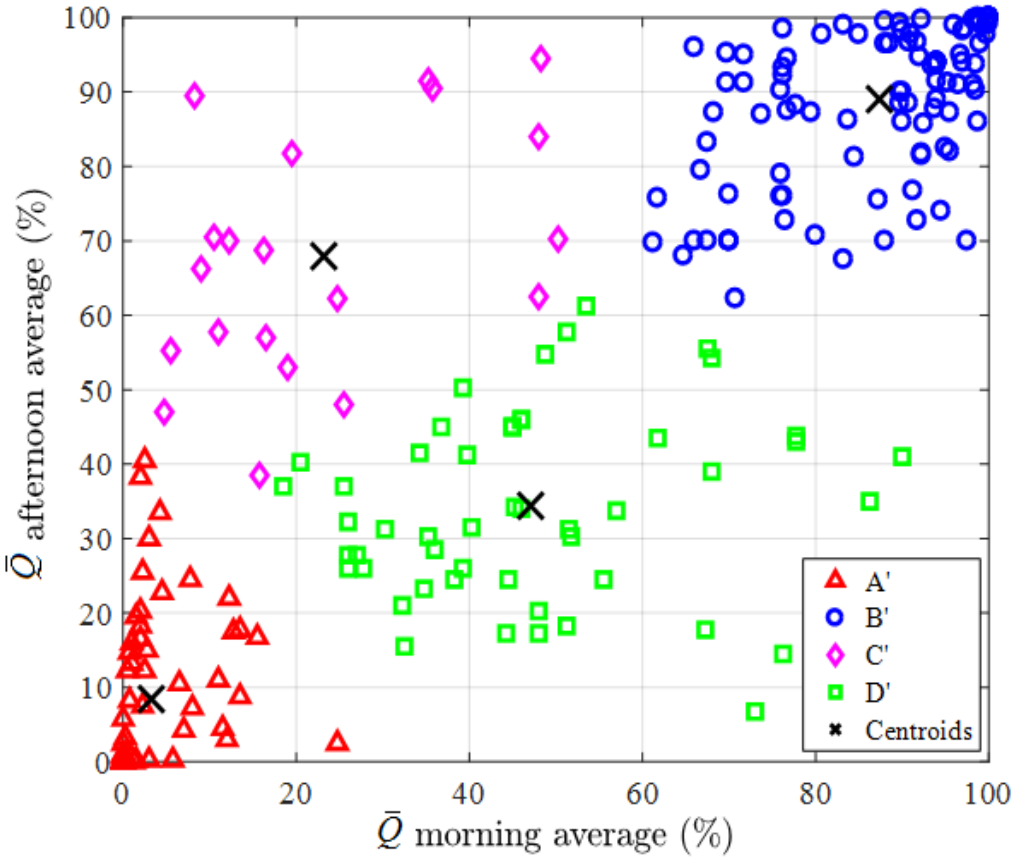


Figure 6.1: Cluster map of Q with daily profiles averaged for morning (9:00-12:00) and afternoon (13:00-16:00).

Table 6.1: Summary of Q clustering.

Class	Cluster	Frequency of days	Proportion	$\overline{SI}_C$	$\overline{SI}_C < 0$
A′	1	63	26%	0.88	0
B′	2	108	44%	0.86	0
C′	3	20	8%	0.46	2
D′	4	52	21%	0.42	4

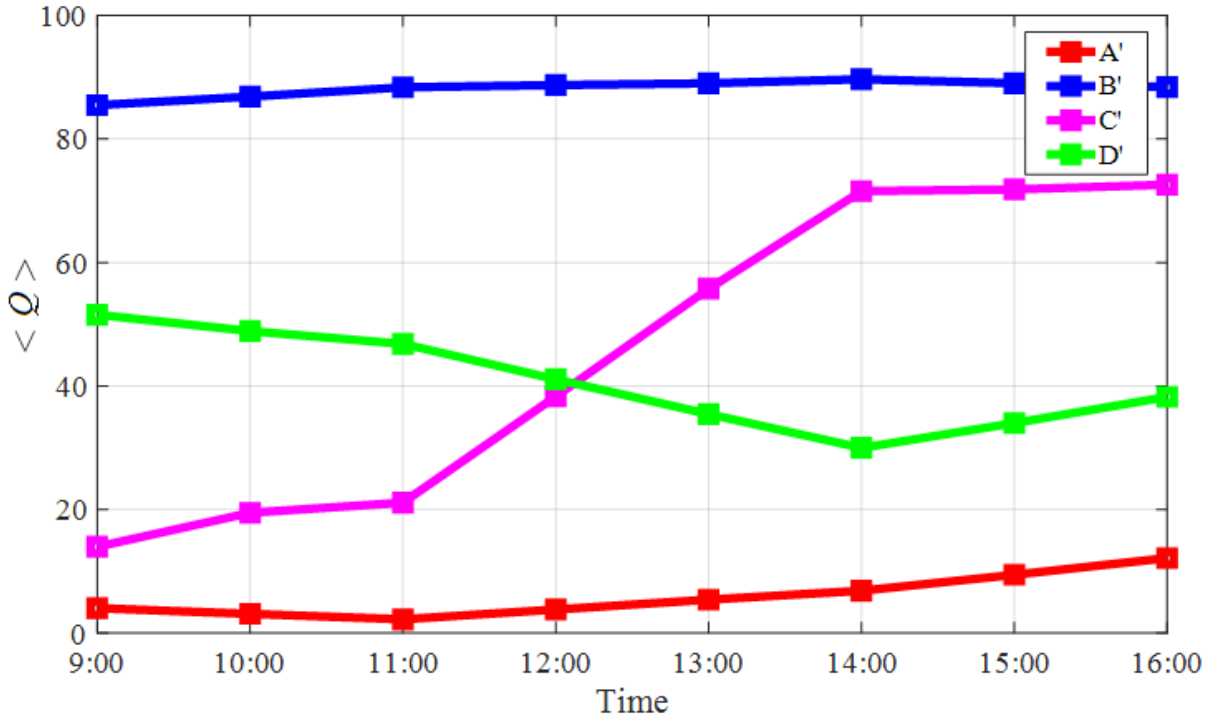


Figure 6.2: Class mean profiles for Q. Class A′ has low levels of Q all day. These correspond to sunny days. Class B′ is the opposite where Q is high throughout the day corresponding to cloudy days. Class C′ has low Q in the morning and high Q in the afternoon, corresponding days that have sunny mornings and cloudy afternoons. Class D′ is the opposite with relatively high Q in the morning and low Q in the afternoon, corresponding to days that have cloudy mornings and sunny afternoons.

6.1.2 Forecasting using Q clustering

The first forecasting method was to use the four classes of Q obtained from clustering daily Q profiles. For a given day, the Q forecast was obtained and the day assigned to one of the Q classes by finding the nearest centroid in the 8-dimensional cluster space. The associated irradiance class was then regarded as the forecast of the irradiance class of that day. The forecast irradiance profile is the class mean profile, with estimated uncertainty given by the class standard deviation. For example, suppose that a given day is forecast to belong to Q Class A$'$. Then the irradiance forecast is that the day belongs to the associated $\bar{B}_n$ Class A. The forecasting results using the clustering of Q are presented in Table 6.2. Each cell in Table 6.2 shows the number of days in a predicted class that were found to be in an actual class, with marginals in the rightmost column and lowest row. Shown in brackets are the number in a cell as a percentage of the total in the rightmost column. For example, the first row shows the 36 days predicted to be in Class A, of which 26 were actually in Class A, 1 in B, 4 in C and 5 in D. Thus 72% of days predicted to be in Class A were actually in Class A, whereas 3% were actually in Class B, 11% in C and 14% in D. The number of correct predictions are in the main diagonal.

Table 6.2: Forecast results using classes of Q on the testing sample of 100 days in 2017. Each cell shows the number of days in a predicted class that were found to be in an actual class, with marginals in the rightmost column and lowest row. Shown in brackets are the number in a cell as a percentage of the total in the rightmost column.

		Actual $\bar{B}_n$ class				
		A	**B**	**C**	**D**	**Total forecast**
Predicted $\bar{B}_n$ class	**A**	26 (72%)	1 (3%)	4 (11%)	5 (14%)	**36**
	B	0 (0%)	23 (66%)	8 (23%)	4 (11%)	**35**
	C	3 (30%)	2 (20%)	5 (50%)	0 (0%)	**10**
	D	2 (11%)	4 (21%)	2 (11%)	11 (58%)	**19**
	Total actual	**31**	**30**	**19**	**20**	**100**

6.1.3 Forecasting using the 50% rule

The second forecasting method uses a simple decision rule, termed the "50% rule" because it assigns a predicted *Q* profile from the NWP to a class depending on the cloud cover averages for the morning and afternoon. This method was chosen as a simple alternative to *k*-means clustering that takes into account the diurnal variation of the $\bar{B}_n$ classes. Forecasting the class of day as Class A, B, C or D using the 50% rule uses a set of decision rules based on Q. The day is partitioned into two parts, i.e. morning and afternoon, where the morning hours are from 9:00 to 12:00 and the afternoon hours are from 13:00 to 16:00. The average Q percentage for the morning and afternoon are taken over the respective periods, and is referred to as Q_{AM} and Q_{PM}, respectively. A set of decision rules presented in Table 6.3 are applied to morning and afternoon Q averages, and the day is assigned to a $\bar{B}_n$ class depending on whether Q_{AM} and Q_{PM} are above or below 50%. The mean profile for the associated $\bar{B}_n$ class is the forecast for the day. The 50% rule produced the forecasting results in Table 6.4.

Table 6.3: Decision rules for morning and afternoon average Q percentage and the associated $\bar{B}_n$ class. Class A (sunny), Class B (cloudy), Class C (sunny morning-cloudy afternoon) and Class D (cloudy morning -sunny afternoon).

Associated $\bar{B}_n$ Class	Q_{AM}	Q_{PM}
A	$\leq 50\%$	$\leq 50\%$
B	$\geq 50\%$	$\geq 50\%$
C	$\leq 50\%$	$> 50\%$
D	$> 50\%$	$\leq 50\%$

Table 6.4 shows that the 50% rule produces a prediction success rate of 63%. The success rate for the individual classes A, B, C and D are 64%, 59%, 63% and 83%, respectively. This method, however, shows increased performance for Classes C and D, and a decrease in performance for Classes A and B.

Table 6.4: Forecasting results using the 50% rule. This table has the same format as Table 6.2.

		Actual $\bar{B}_n$ class				
		A	**B**	**C**	**D**	**Total forecast**
Predicted $\bar{B}_n$ class		30	4	4	9	**47**
	A	(64%)	(9%)	(9%)	(19%)	
		0	23	10	6	**39**
	B	(0%)	(59%)	(26%)	(15%)	
		1	2	5	0	**8**
	C	(13%)	(25%)	(63%)	(0%)	
		0	1	0	5	**6**
	D	(0%)	(17%)	(0%)	(83%)	
	Total actual	**31**	**30**	**19**	**20**	**100**

Table 6.5 shows the average RMSE values, computed using equation 3.2, for each class using the respective forecasting method. Overall, the average RMSE per class are fairly similar using the two forecasting methods, with smaller RMSE for Classes A and B using Q clustering, and smaller RMSE for Classes C and D using the 50% rule. The largest difference was in Class A.

Table 6.5: Average RMSE for each class using the two forecasting methods for day ahead forecasts of $\bar{B}_n$.

Class	Q clustering	50% rule
A	0.20	0.27
B	0.22	0.24
C	0.32	0.29
D	0.34	0.29

6.2 Hourly forecasts of $\bar{B}_n$ and $\bar{D}_n$

6.2.1 Forecasting using Persistence of the Class Trend

In addition to day forecasts, intra-day forecasts of $\bar{B}_n$ and $\bar{D}_n$ were also investigated. The $\bar{B}_n$ and $\bar{D}_n$ quantities were forecast separately since they each have a set of class mean profiles that are used for the intra-day forecasting method. The method used to forecast hourly values of $\bar{B}_n$ and $\bar{D}_n$ used classes that were established through clustering of $\bar{B}_n$. As seen in Chapter 5, each $\bar{B}_n$ class has an associated $\bar{D}_n$ mean profile that characterizes the class in terms of the diffuse irradiance levels. To produce forecasts at hourly intervals, these mean profiles were used to forecast $\bar{B}_n$ and $\bar{D}_n$ for the next hour based on how far away $\bar{B}_n$ and $\bar{D}_n$ are from the class mean profile in the current hour. The method called "Persistence of the Class Trend" (PCT) uses the mean profile of the class as a reference, together with the current actual $\bar{B}_n$ and $\bar{D}_n$ values, to forecast $\bar{B}_n$ and $\bar{D}_n$ for the next hour, according to the equations 6.1 and 6.2. The PCT forecasting method is illustrated in Figure 6.3.

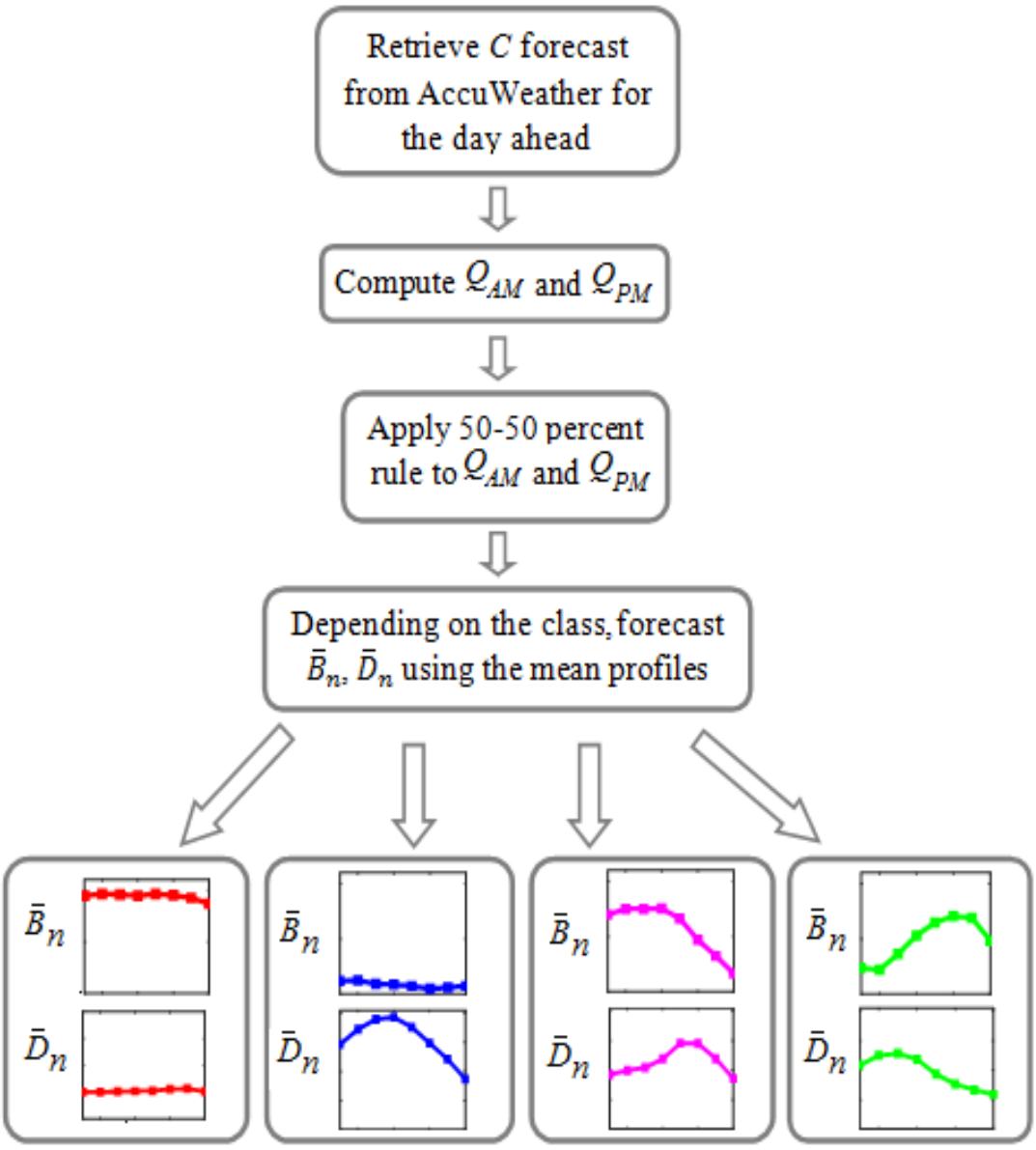


Figure 6.3: Description of the PCT method. The Q forecast for the day is retrieved from AccuWeather and Q_{AM} and Q_{PM} are computed. The 50% rule is applied to forecast the class of day and thereafter the mean profiles for that class are used to forecast hourly values of $\bar{B}_n$ and $\bar{D}_n$.

First, the Q forecast is obtained from AccuWeather the start of the day. Then Q is averaged for the morning and afternoon and the 50% rule from Table 6.3 is applied to forecast the day as Class A, B, C or D. Then, once the class of day is established, the PCT method is applied using the corresponding mean profile for $\bar{B}_n$ and $\bar{D}_n$ for that specific class.

An example of a Q forecast from AccuWeather is given in Figure 6.4. This particular profile is

typical of a sunny day in Durban.

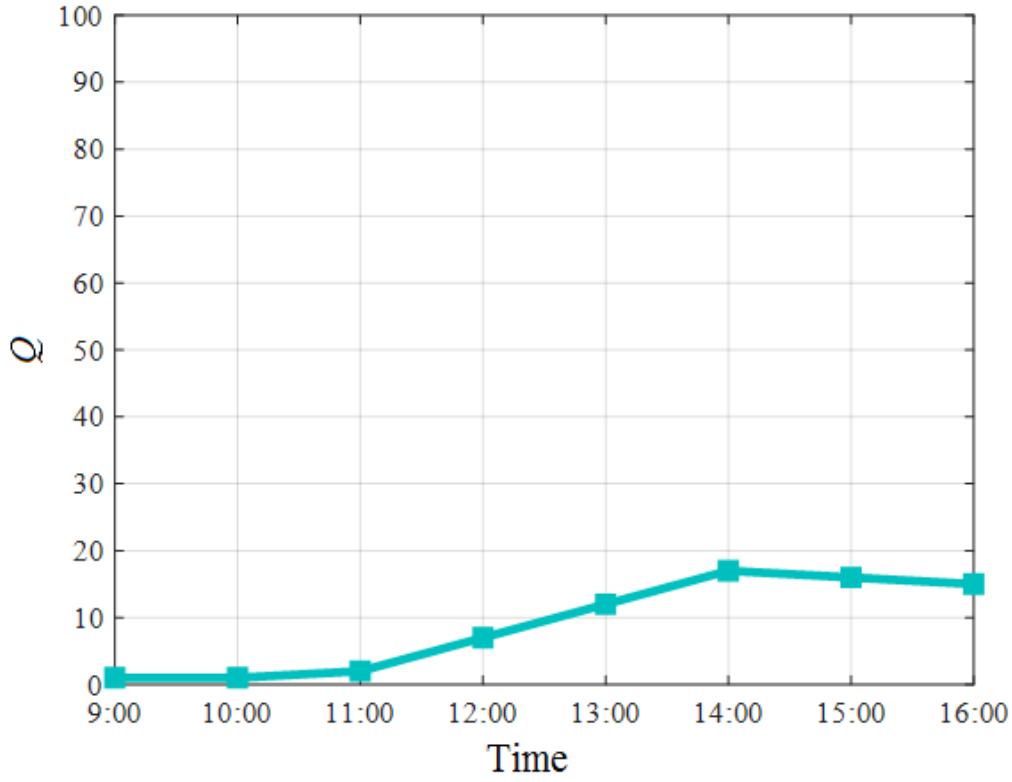


Figure 6.4: An example of a Q forecast obtained from AccuWeather that is required for the PCT method. This Q profile was the forecast for 29 July 2017. This profile is typical of a sunny day in Durban, where Q is low throughout the day.

Using the class mean profile, the hour-ahead forecast of $\bar{B_n}$ denoted as $B_{n_{forecast}}$ is given by

$$B_{n_{forecast}} = \frac{B_{n_{actual}} \times M_{t+1}}{M_t}, \tag{6.1}$$

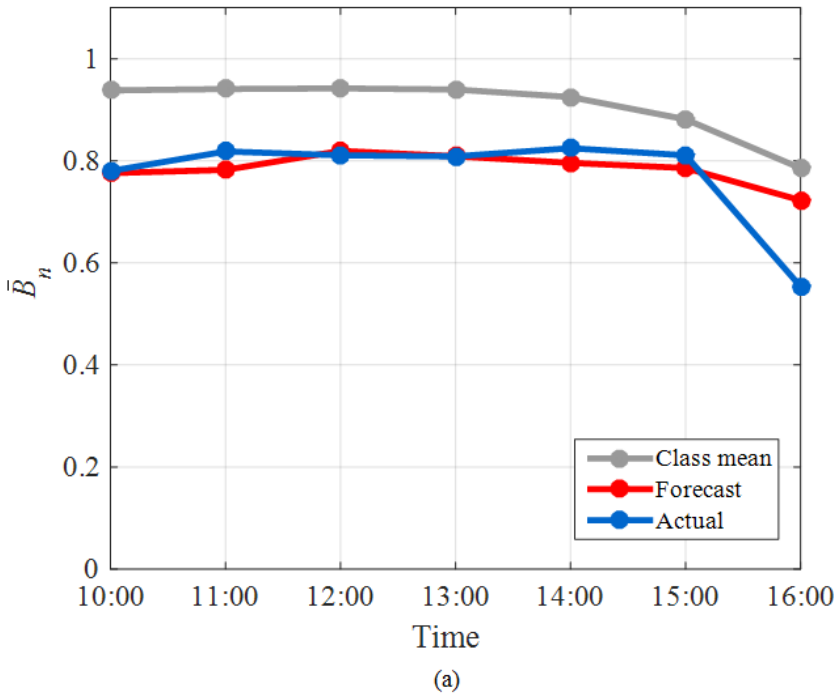
where $B_{n_{actual}}$ is the actual $\bar{B_n}$ value, M_t is the mean of the $\bar{B_n}$ class at hour t and M_{t+1} is the mean of the $\bar{B_n}$ class at hour $t+1$. Similarly, the hour-ahead forecast for $\bar{D_n}$, denoted as $D_{n_{forecast}}$ is given by

$$D_{n_{forecast}} = \frac{D_{n_{actual}} \times M_{t+1}}{M_t}, \tag{6.2}$$

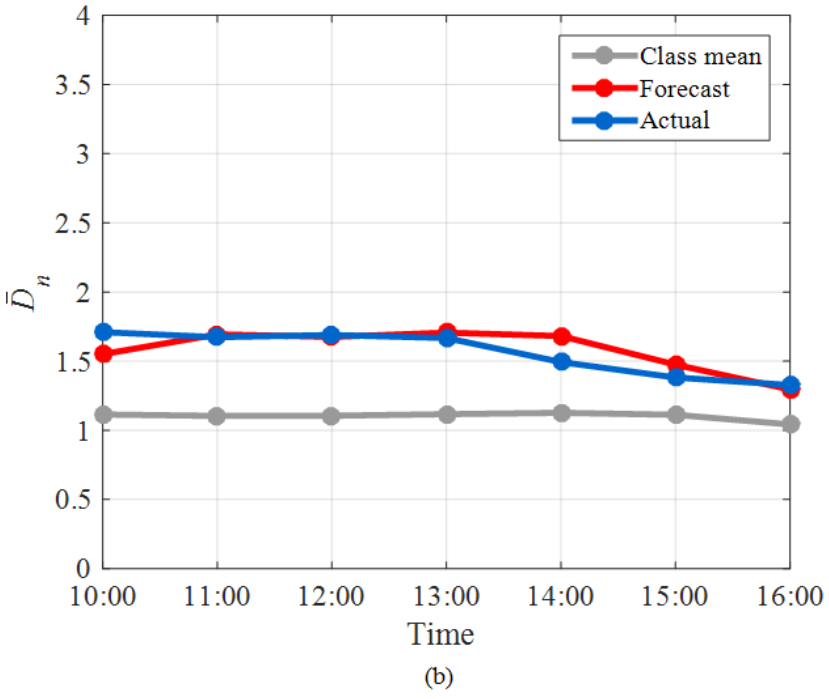
where $D_{n_{actual}}$ is the actual $\bar{D_n}$ and where in this case, M_t is the mean of the $\bar{D_n}$ class at hour t and

M_{t+1} is the mean of the $\bar{D_n}$ class at hour $t+1$. The a first actual irradiance value is required by the PCT method to estimate how far the actual value is from the mean profile. As a result, hourly forecasts of $\bar{B_n}$ and $\bar{D_n}$ start an hourly later.

Figures 6.5-6.8 (a) and (b) illustrate the PCT method for hourly forecasts of $\bar{B_n}$ and $\bar{D_n}$. Examples of each class are shown with the actual and forecasted profiles for $\bar{B_n}$ and $\bar{D_n}$. The mean profile of the class was also included to illustrate the effect of the PCT method.

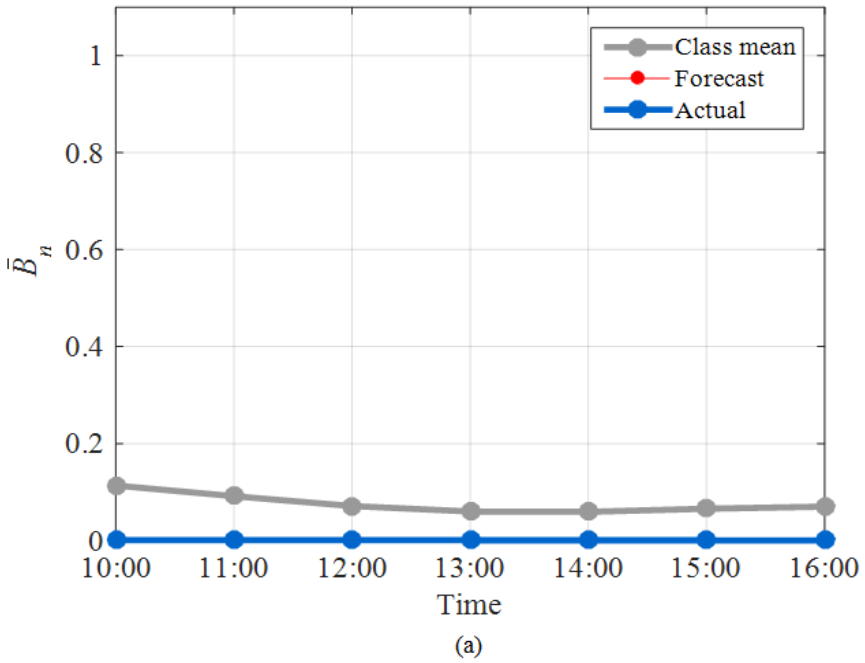


(a)

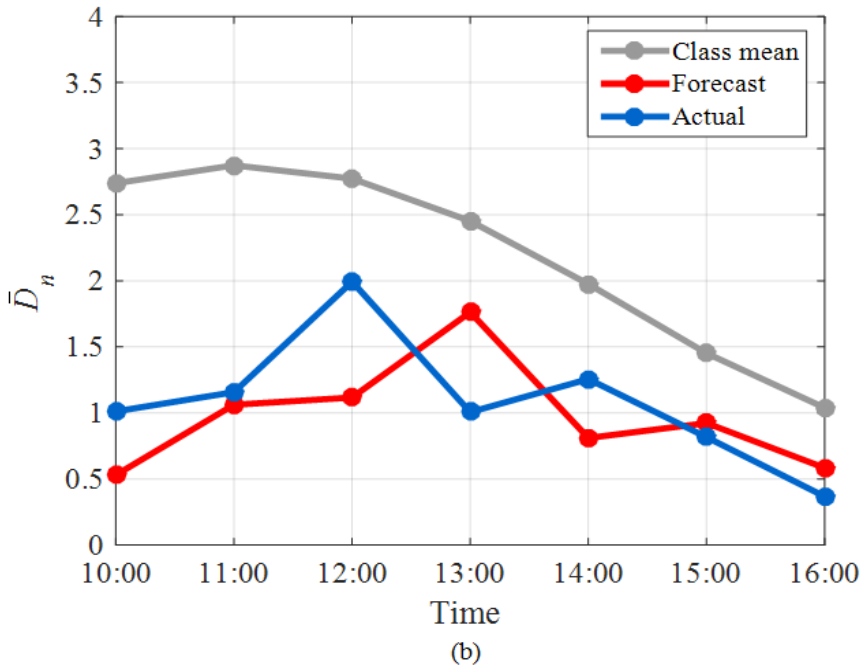


(b)

Figure 6.5: Example of Class A using the PCT method for 2 April 2017. Hourly forecast of (a) $\bar{B}_n$ and (b) $\bar{D}_n$.

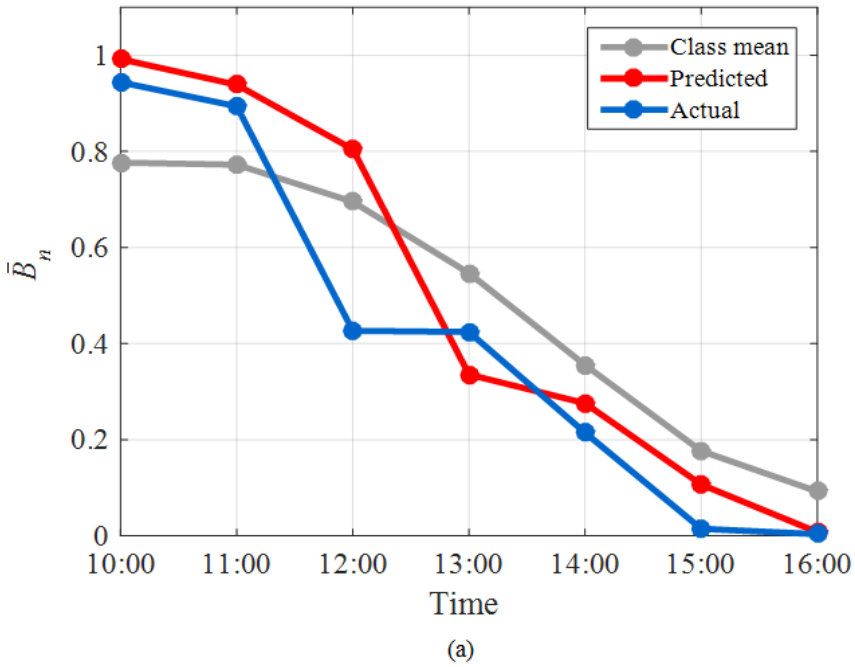


(a)

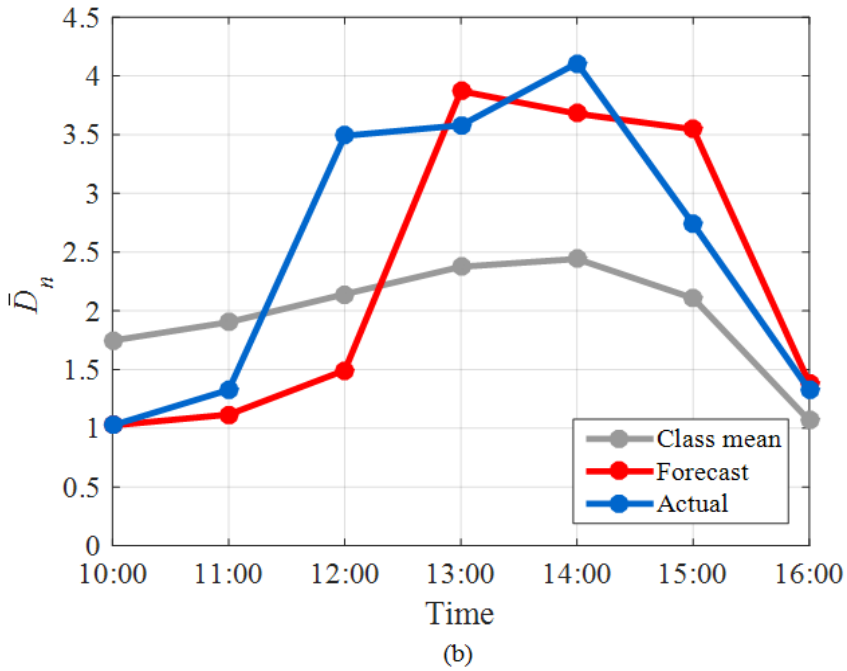


(b)

Figure 6.6: Example of Class B using the PCT method for 14 April 2017. Hourly forecast of (a) $\bar{B}_n$ and (b) $\bar{D}_n$.

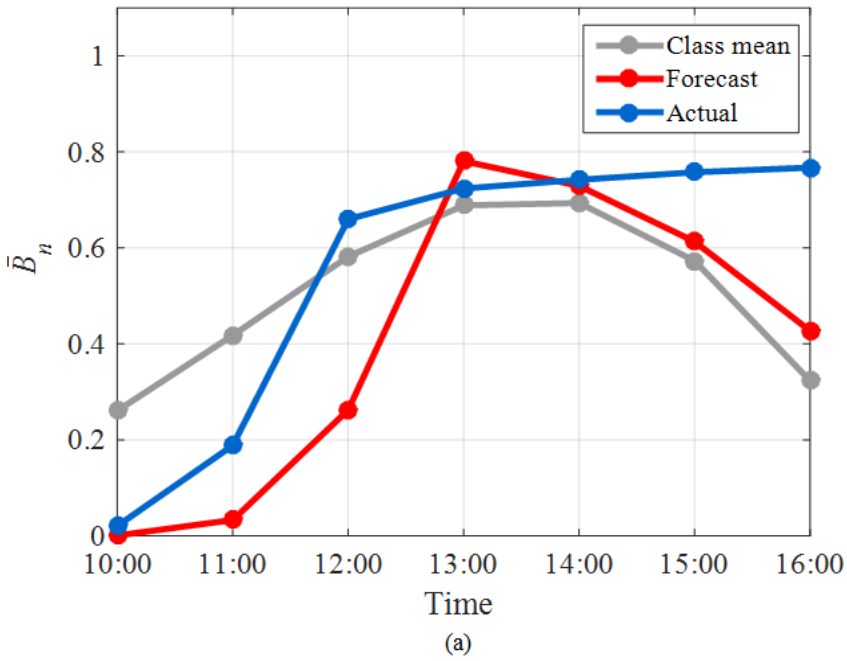


(a)



(b)

Figure 6.7: Example of Class C using the PCT method for 18 March 2017. Hourly forecast of (a) $\bar{B}_n$ and (b) $\bar{D}_n$.



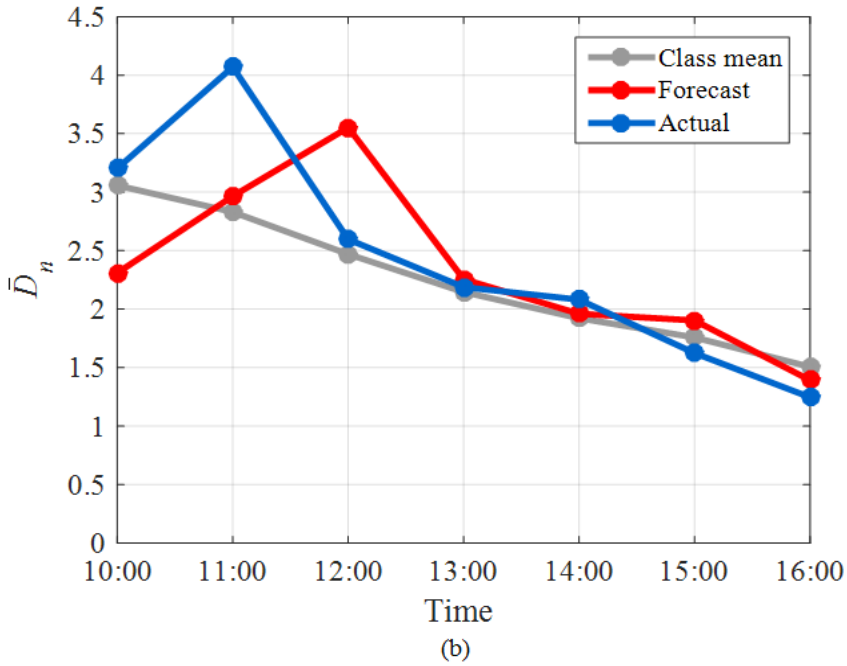


Figure 6.8: Example of Class D using the PCT method for 5 April 2017. Hourly forecast of (a) $\bar{B}_n$ and (b) $\bar{D}_n$.

6.2.2 Forecast error using the PCT method

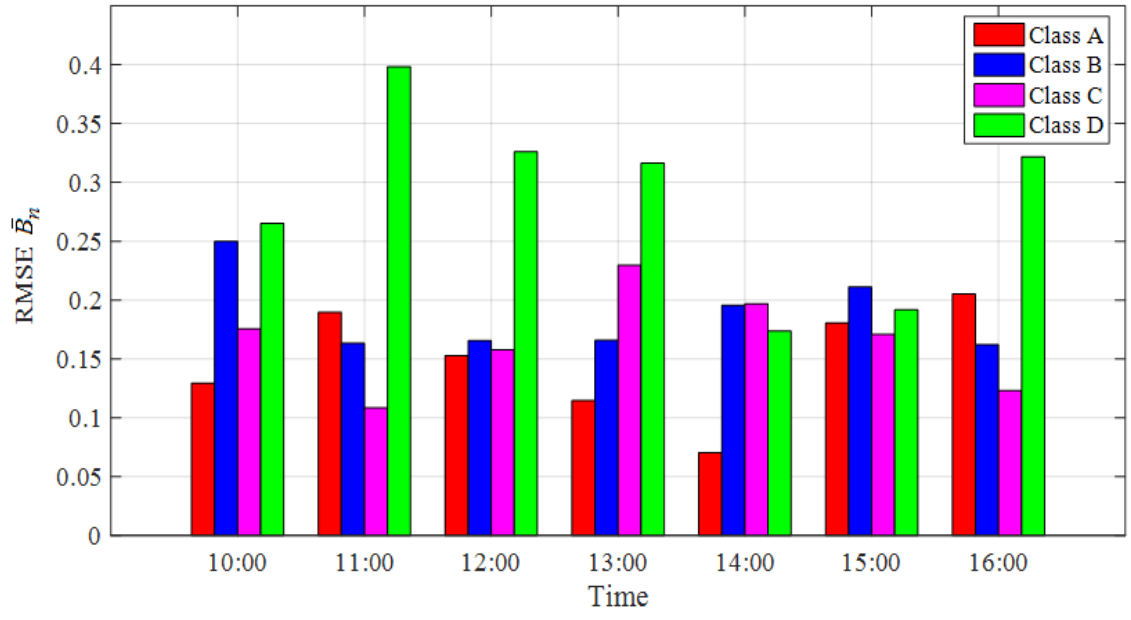
Performance of the PCT forecasting method is quantified using the RMSE. This is the most commonly-used error metric for comparing how close a forecast is to the actual. For each class, the RMSEs for $\bar{B}_n$ and $\bar{D}_n$ for all the days in that hour are presented in Figure 6.9. In addition, the variance for $\bar{B}_n$ (denoted as σ_{B_n}) and $\bar{D}_n$ (denoted as σ_{D_n}) in each hour is shown in Figure 6.10.

6.2.3 Comparison to Persistence

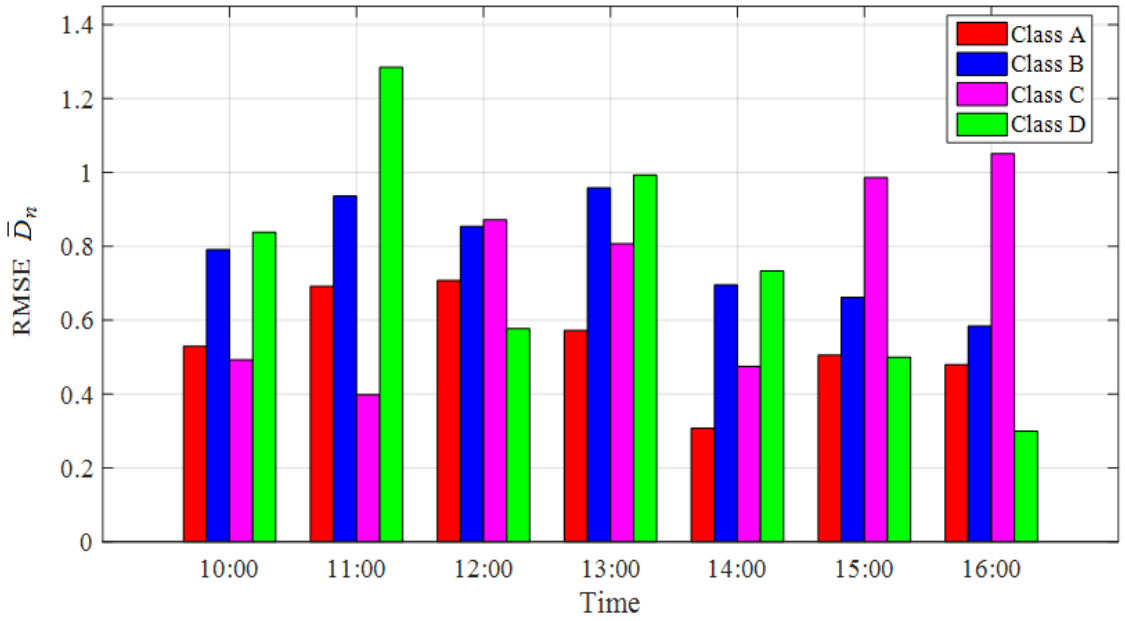
To establish the performance of a forecasting method, a comparison is made using the Persistence forecasting method. The Persistence method was applied to the same set of test days as the PCT method and the RMSE for each hour was computed. Tables 6.6-6.9 present the percentage improvement in the RMSE in each hour for Classes A to D, respectively. In each table the percentage improvement in $\bar{B}_n$ and $\bar{D}_n$ for the hour are in the second and third columns. From the total number of days in each class that were tested, not all forecasts using the PCT method showed an improvement over Persistence. This is expected since Persistence is difficult to improve upon when sky conditions remain fairly static i.e. sunny and cloudy days. The percentages presented in these tables are for the days where the PCT method did show an improvement over Persistence. Listed in the last two columns are the number of days (out of the class total) for $\bar{B}_n$ and $\bar{D}_n$ that were found to have shown improvements over Persistence, and are denoted as N_{B_n} and N_{D_n}.

For Class A, the PCT method showed most improvement in $\bar{B}_n$ during the afternoon. The improvement in $\bar{D}_n$ was relatively low ($< 5\%$) for all hours with the exception of hour 16:00. Classes B, C and D show higher improvements in $\bar{B}_n$ and $\bar{D}_n$, but the sample sizes for Classes C and D are much smaller and this increases the percentage quite substantially. The average improvement for all hours in each class is given in the last row of each table. Overall, the average improvement of the PCT method over Persistence for all classes was found to be 22.2% for $\bar{B}_n$ and $\bar{D}_n$.

For all classes, for the remaining days for which Persistence Forecasts performed better than the PCT method, the average decrease in Persistence RMSE was found to be 35% and 25% for $\bar{B}_n$ and $\bar{D}_n$, respectively. For $\bar{B}_n$ this is slightly higher that that of McCandless et al. (2015), but for $\bar{D}_n$ it is relatively similar.

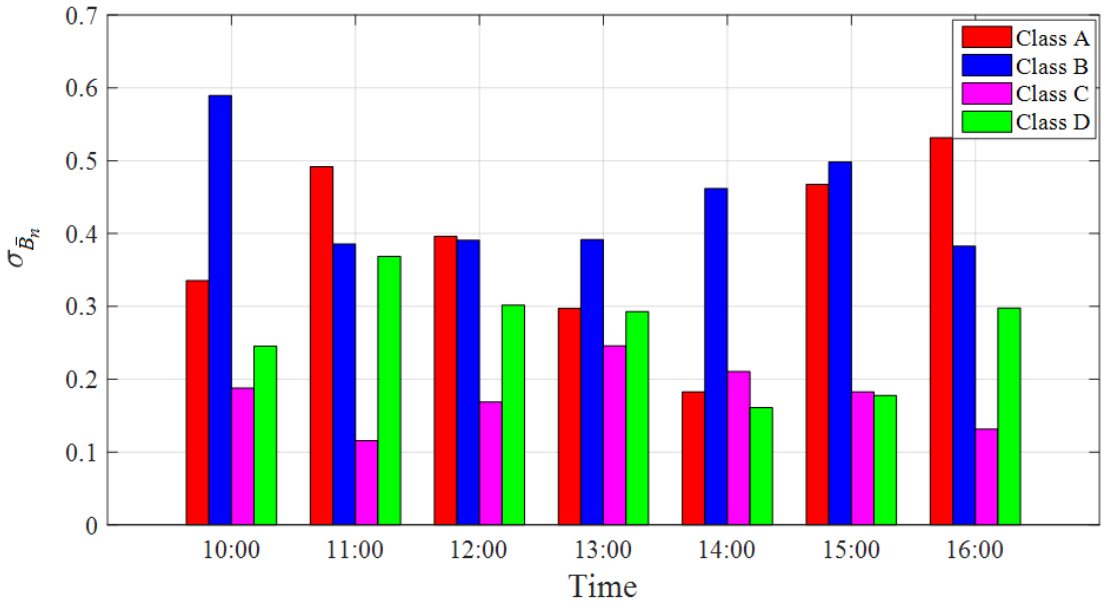


(a)

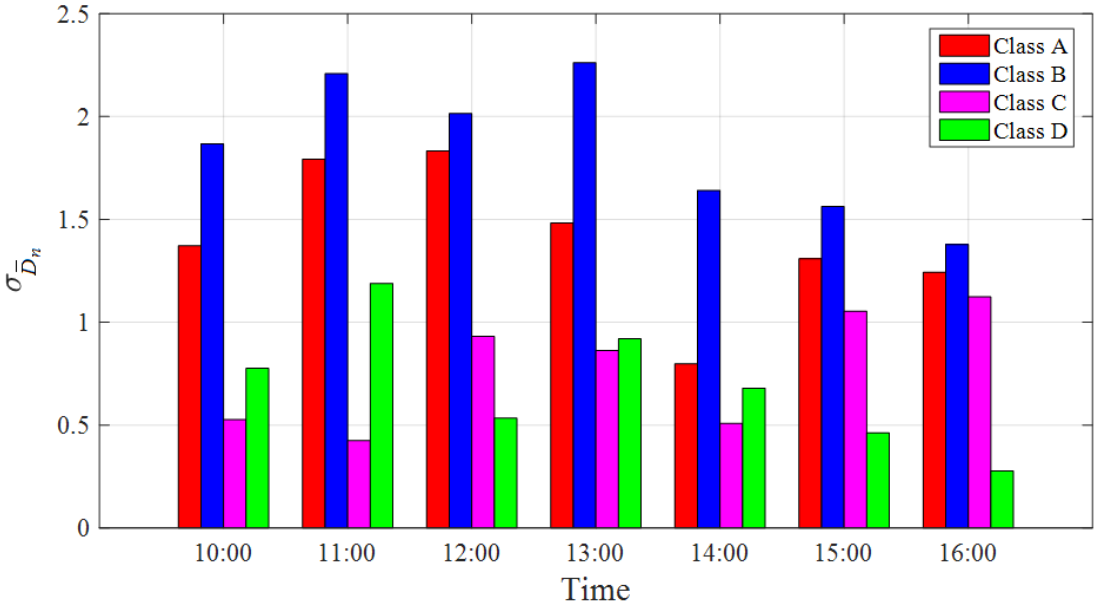


(b)

Figure 6.9: RMSE for Classes A-D using the PCT method. (a) RMSE for each hour of $\bar{B}_n$ (b) RMSE for each hour of $\bar{D}_n$.



(a)



(b)

Figure 6.10: Variance for Classes A-D using the PCT method. (a) variance for each hour of $\bar{B}_n$ (b) Variance for each hour of $\bar{D}_n$.

Table 6.6: Class A percentage improvement in RMSE over Persistence using the PCT forecasting method, for the set of days indicated the in last two columns. Class A has a total of 47 days.

Hour	**Improvement in B_n RMSE**	**Improvement in D_n RMSE**	N_{B_n}	N_{D_n}
10:00	1.3%	2.6%	25	20
11:00	0.43%	1.7%	21	25
12:00	0.2%	0.1%	29	18
13:00	2.6%	3.1%	28	26
14:00	6.7%	1.1%	20	20
15:00	10.3%	3.5%	22	15
16:00	17.4%	14.8%	23	29
Average	**5.6%**	**3.8%**		

Table 6.7: Class B percentage improvement in RMSE over Persistence using the PCT forecasting method, for the set of days indicated the in last two columns. Class B has a total of 39 days.

Hour	**Improvement in B_n RMSE**	**Improvement in D_n RMSE**	N_{B_n}	N_{D_n}
10:00	13.4%	21.7%	14	27
11:00	33.2%	7.4%	14	12
12:00	48.9%	7.3%	17	16
13:00	26.2%	13.6%	17	18
14:00	1.5%	42.8%	20	14
15:00	14.2%	52.6%	10	20
16:00	1.8%	50.6%	5	30
Average	**19.9%**	**28.0%**		

Table 6.8: Class C percentage improvement in RMSE over Persistence using PCT the forecasting method, for the set of days indicated the in last two columns. Class C has a total of 8 days.

Hour	**Improvement in B_n RMSE**	**Improvement in D_n RMSE**	N_{B_n}	N_{D_n}
10:00	8.5%	7.7%	5	5
11:00	0.3%	23.6%	5	5
12:00	42.3%	13.6%	5	5
13:00	76.5%	0%	2	1
14:00	60.6%	16.9%	4	4
15:00	44.1%	29.6%	7	5
16:00	89.9%	89.1%	5	3
Average	**43.0%**	**25.0%**		

Table 6.9: Class D percentage improvement in RMSE over Persistence using PCT the forecasting method, for the set of days indicated the in last two columns. Class D has a total of 6 days.

Hour	**Improvement in B_n RMSE**	**Improvement in D_n RMSE**	N_{B_n}	N_{D_n}
10:00	2.1%	15.3%	5	3
11:00	41.6%	24.1%	3	2
12:00	8.0%	40.3%	6	5
13:00	14.7%	24.5%	5	4
14:00	1.7%	24.5%	5	4
15:00	30.9%	40.9%	2	2
16:00	44.7%	53.2%	2	3
Average	**20.5%**	**31.8%**		

Chapter 7

Discussion

A discussion of the results from the clustering analysis and forecasting are presented in this chapter. A comparison of the clustering results for the different variables, and a comparison of the forecasting methods using classes are discussed. The chapter concludes with an overall summary of the classification and forecasting results for Durban.

7.1 Classification of irradiance profiles

7.1.1 Minute-resolution irradiance profiles

One of the aims of this research was to use clustering for classification of irradiance patterns. Classification of the solar irradiance patterns in Durban was obtained by clustering of minute-resolution B_n profiles. Due to the large number of dimensions i.e. 481 in the minute-resolution profiles, PCA was applied as a pre-processing technique. As indicated by the scree plot in Figure 5.3, the first 8 components explain more than 90% of the variance in the data, therefore reducing the number of dimensions to 8 from the original 481. Table 7.1 lists the individual percentage and cumulative percentages of the first 8 components. This demonstrates that for Durban the first 8 components are sufficient to describe 90% of the variance in all dimensions.

The solution of four clusters produced a relatively high SI_{TOT} value i.e. 0.64, in keeping with the benchmark of a good silhouette value suggested by Lletí et al. (2004). Furthermore, this solution produced 5% of days with a negative SI which indicates that 95% of the days in Durban are well represented by these four clusters. The clustered B_n profiles have a set of associated D_n profiles.

Table 7.1: Individual and cumulative percentage variance for the first 8 Principal Components for B_n minute-resolution profiles.

Principal component	Individual percentage variance	Cumulative percentage variance
1	72	72
2	8	80
3	4	84
4	2	86
5	1	87
6	1	88
7	0.8	89
8	0.8	90

The mean profiles in Figures 5.5 and 5.6 characterize the B_n and D_n classes and describe the irradiance levels for the day.

Although the present work focused on forecasting for the day-ahead, Figures 5.12 (a) and (b) may be used in future work for developing models for forecasting the irradiance class for more than one day ahead. Figure 5.12 (a) shows the frequency of classes occurring in sets. Mostly, only Classes A and B occur for more than three consecutive days. Figure 5.12 (b) shows the frequency of next day class occurrences. If forecasting for more than one day ahead is pursued, this can assist in deciding the most appropriate forecasting tools. For example, if the current day is a Class A, the next day is most likely to be a Class A and therefore forecasting tools such as satellite and ground-based imagers may not be required to track cloud motion and instead statistical techniques could be sufficient. On the other hand, if the current day is a Class D, the next day could either be either a Class A, B, or C. In this situation it would be necessary to have satellite and ground-based cloud imagery to monitor large cloud masses and track individual clouds.

7.1.2 Hourly-resolution irradiance profiles

In order to match the hourly-resolution of the NWP cloud cover, and to investigate whether the same diurnal patterns can be regained with lower temporal resolution data, hourly-resolution profiles, $\bar{B}_n$, were clustered. The class mean profiles in Figure 5.15 show the same diurnal patterns were indeed

regained. This high degree of similarity indicates that the sub-hourly temporal structure in B_n did not result in any significant difference in clustering compared with the temporal structure in the hourly average. These four $\bar{B_n}$ and $\bar{D_n}$ classes successfully characterize the solar irradiance patterns in Durban.

As mentioned in Section 5.3, $\bar{B_n}$ profiles presented in Figure 5.15 show a difference in the strength of the irradiance levels in the sunny regions for Classes C and D as compared to Class A. More specifically, the values of $\bar{B_n}$ in the sunny regions of Classes C and D are significantly lower ($0.6 < \bar{B_n} < 0.8$) than that of Class A (> 0.9). In Durban, low clouds occur more frequently in the late afternoon or evening, as well as in the early morning hours (South African Weather Service, 2010). As described by Wood (2012), stratocumulus clouds exhibit strong diurnal modulation largely due to the diurnal cycle of solar insolation and consequently absorption of solar radiation during the daytime in the upper regions of the cloud. This results in a suppression of the total radiative driving, resulting in weaker circulation during daytime than at night and a less efficient coupling of the clouds with the surface moisture supply. Because of this, the maximum coverage of stratocumulus tends to be during the early hours of the morning. Eastman and Warren (2013) state that low clouds respond to the day-night cycle of solar flux. During the day, solar heating of the surface results in convection and cumuliform cloud formation due to destabilization of the atmospheric boundary layer (i.e. the lowest part of atmosphere that is closest to the Earth's surface). At night the boundary layer cools, resulting in condensation and the formation of stratiform clouds. Both cumuliform and stratiform clouds fall into the category of low cloud types. This may explain the possibility of Classes C and D having reduced morning and afternoon average $\bar{B_n}$ levels. More specifically, for Class C days, if there is stratocumulus or cumulus cloud present in the early morning, they may not completely dissipate by 9:00 and so the morning average $\bar{B_n}$ level will be lower that than of Class A. Similarly, for Class D days, if there is stratocumulus cloud present in the late afternoon, there may be presence of their formation during the early afternoon (14:00-16:00) that results in a lower afternoon average $\bar{B_n}$, as compared to the Class A.

The possibility of the presence of cloud in the sunny regions of Classes C and D may be observed by the difference in the morning $\bar{D_n}$ levels of Classes C and D from Class A. More specifically, Class A has a morning $\bar{D_n}$ average of about 1, whereas Classes C and D have a morning $\bar{D_n}$ average that exceeds 1.5 despite their being classified as a "sunny" morning. For Class D, the morning average $\bar{D_n}$ is more than double that of Class A. According to Haurwitz (1948), low clouds reduce the in-

solation by approximately 65%. This implies that the diffuse irradiance may be almost doubled in the presence of low clouds. Therefore, this could further explain the possibility of the occurrence of low cloud in Class D, where the morning $\bar{D_n}$ average is more that twice the level of Class A.

In Figure 5.16, the $\bar{D_n}$ profile for Class B exceeds 3.5 during midday. According to Miller (1981), as cloud cover fraction increases from about 0.3 to 0.7, the diffuse irradiance increases by about half, from 130 W/m^2 to 200 W/m^2. Furthermore, the midday flux density or diffuse radiation ranges up to 200-300 W/m^2. This could be an explanation for the $\bar{D_n}$ Class B trend that has been observed, i.e. increasing trend towards midday where it peaks and thereafter decreases toward the afternoon.

Durban frequently experiences cold fronts, which develop as closed low-pressure cells in the Western Atlantic and move across Southern Africa in a west-northwest to an east-southeast direction (Nel, 2009). As shown in Figure 5.12 (a), they most frequently last for 2 consecutive days, and occasionally for 3. Cold fronts exceeding 3 consecutive days are rare for Durban. It is therefore expected that days in Classes C and D may be a transition between days in Classes A nd B. More specifically, the cloudy afternoon of Class C suggests a start of a 2 day cold front, and the cloudy morning of Class C suggests an end of a cold front. Since Figure 5.15 suggests that Classes C and D contain days that undergo a transition from a sunny morning to cloudy afternoon and vice-versa, this could explain the distinctly lower $\bar{B_n}$ in the sunny regions. A possible reason is that lower $\bar{B_n}$ is observed since in order to transition from a sunny morning to a cloudy afternoon it is expected that there will be some clouds present in the morning, where their formation progresses toward midday and becomes more distinct in the afternoon. Similarly for Class D, the day transitions from a cloudy morning to a sunny afternoon, where clouds tend to dissipate during midday and results in a sunny afternoon. However, the maximum $\bar{B_n}$ in the afternoon for Class D will be lower than that of Class A due to the presence of some residual cloud cover.

Another feature of $\bar{D_n}$ seen in Figure 5.16 is that $\bar{D_n}$ for Classes B, C and D have a decreasing trend toward the afternoon. $\bar{D_n}$ for Class B reaches a maximum around midday (11:00-13:00), decreases substantially after 13:00, and reaches to an almost clear sky level at 16:00. During this time, the $\bar{B_n}$ profile of Class B remains below 0.2 throughout the day with little fluctuation in the average trend. It is possible that during cloudy days, $\bar{D_n}$ weakens in the afternoon (i.e. after 13:00) and it may also be weak in the early morning. However, the asymmetry of the time interval over which the clustering was applied prevents inspection of $\bar{D_n}$ before 9:00 (for example at 8:00 or at 7:00).

From Figure 5.14 it can be seen that Class C has very few days that have mornings as sunny as the mornings in Class A i.e. $\bar{B}_n \approx 1$. The cluster map suggests that Class C days that have a $\bar{B}_n$ morning average in the range 0.6-0.8, tend to have $\bar{B}_n$ afternoon averages less than 0.4. Similarly, only a few days in Class D reach the same $\bar{B}_n$ afternoon average as that of a Class A. According to the cluster map, almost all days in Class D will reach a maximum of about 0.8 in the afternoon. The cluster map therefore confirms that Classes C and D do not reach the same maximum levels in their sunny regimes as does Class A.

As observed in Figure 5.11, Durban is dominated by sunny days (Class A), followed by cloudy days (Class B). Days that have sunny mornings and cloudy afternoons (Class C) are prevalent during the months of August and September. Days that start off cloudy in the morning and become sunny in the afternoon (Class D) occur mostly during October. Figure 7.1 shows the yearly evolution of the $\bar{B}_n$ classes where Clusters 1, 2, 3 and 4 denote Classes A, B, C and D, respectively. This further illustrates how sunny and cloudy days (Classes A and B) dominate among the 4 irradiance patterns in Durban.

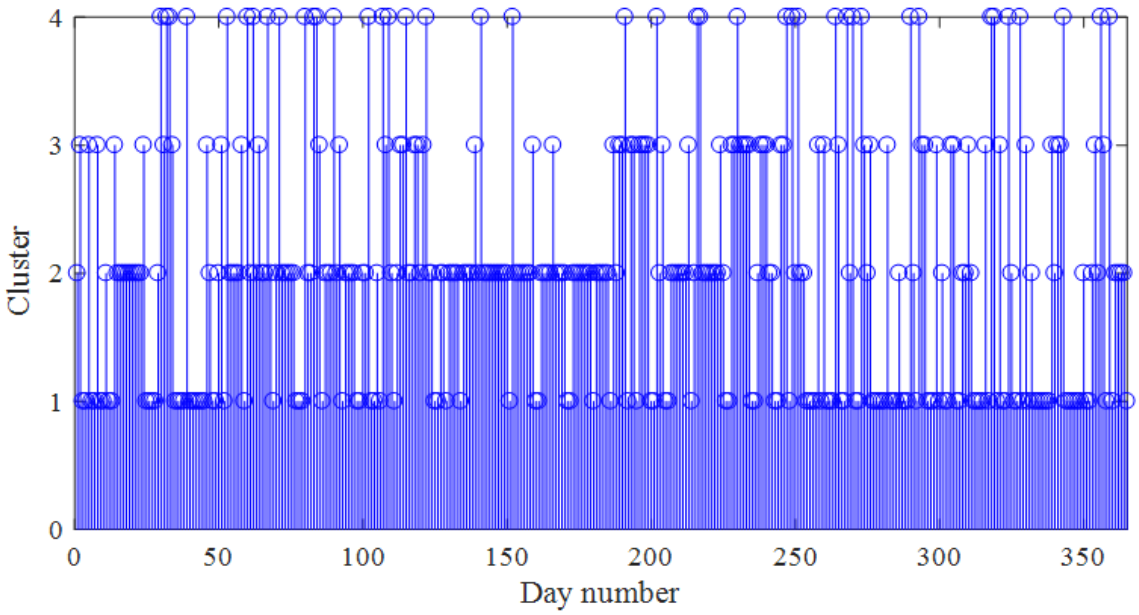


Figure 7.1: Sequence of irradiance classes where clusters 1, 2, 3 and 4 denote Class A, B, C and D, respectively. Classes A and B are dominant throughout the year.

The $\bar{B}_n$ classes established for Durban are very similar to four of the five classes obtained for Reunion Island by Jeanty et al. (2013) and Badosa et al. (2013). The fifth class that resulted from

the clustering by Jeanty et al. (2013) consist of days that are sunny until mid-morning (9:00-9:30) and cloudy thereafter, which was termed "intermittent bad days". These types of days were found to be dominant from November to January (i.e. during summer). Jeanty et al. (2013) suggested that this could be an effect of the land breeze combined with trade winds that could produce orographic clouds (clouds that develop in response to the forced lifting of air by the local topography for example, mountains). As outlined by Réchou et al. (2014), there is the formation of deep convective systems caused by high relative humidity and weak trade winds that are present during the island's summer season, which may be another explanation for the occurrence of intermittent bad days. The orographic clouds produced by the complex terrain of Reunion Island coupled with specific meteorological conditions (high relative humidity and weak trade winds) may be responsible for producing the fifth irradiance class that is dominant during the island's summer season, and which does not appear to be present in Durban's classification system.

As described by Badosa et al. (2015), the advected trade cumuli and large-scale cloud systems are responsible for only a part of the cloudiness affecting the island, since in most cases clouds are formed locally by convection as a result of the interaction between synoptic wind, local thermal winds and the orography. This local cloud formation has a pronounced diurnal cycle as it is driven by the combined effects of trade winds and thermal winds.

For Durban, the seasonal distribution of low cloud is approximately similar to the distribution of rainfall (South African Weather Service, 2010). As described by Dedekind et al. (2016), Southern Africa receives most of its rainfall during the period October to March. It is therefore expected that the summer months are the cloudiest of the year, which is consistent with the high occurrence of Class B days in Durban during summer. This is clearly observed in Figure 5.11, where the months of October to March do indeed show the largest percentage of Class B days, as compared to the months of April to September.

The $\bar{B}_n$ classes established for Durban were also similar to the three classes obtained by McCandless et al. (2014), of which the sunny (Class A) and cloudy days (Class B) formed two distinct classes whereas the partly cloudy days in the McCandless et al. (2014) classification were subdivided here by diurnal pattern into Classes C and D. Even though the work by Jeanty et al. (2013) and Badosa et al. (2013) served as a basis for much of this study, they did not use the classes established through clustering for forecasting. Furthermore, in contrast with Badosa et al. (2013) neither studies by Jeanty et al. (2013) or McCandless et al. (2014) use a combination of irradiance-derived variables

for clustering such as $\{\bar{B}_n, V_B\}$ or $\{\bar{B}_n, \bar{D}_n\}$ and instead only cluster with a single variable.

Results from the clustering of V_B on its own did not show a stronger clustering pattern than for $\bar{B}_n$. Clustering V_B by itself results in the V_B classes having a mixture of different $\bar{B}_n$ class days. For example, both sunny and cloudy days have low variability, and when clustered using V_B only they fall into the same class even though they have distinctly different $\bar{B}_n$ profiles, which was expected. Considering that one of the aims of the present work is to distinguish between sky conditions, V_B was not the best clustering variable to achieve this.

In order to investigate the effect of variability, $\{\bar{B}_n, V_B\}$ were clustered together as a 16-dimensional set. It was found that clustering of $\{\bar{B}_n, V_B\}$ yielded a solution that was driven more strongly by the $\bar{B}_n$ quantity, and hence did not necessarily result in a stronger clustering solution despite the additional information of variability.

Similar to the $\{\bar{B}_n, V_B\}$ combination, the purpose of the combination $\{\bar{B}_n, \bar{D}_n\}$ was to investigate whether the clustering produced a better solution when combined with diffuse irradiance. Figure 5.29 shows that Class A is compact and well separated and Class B is less compact but still relatively separated. Classes C and D, however, have low compactness and are no longer distinctly separated. This is more evident in the class mean profiles of $\bar{B}_n$ in Figure 5.31, where the diurnal patterns in Classes C and D are significantly suppressed. Therefore, the clustering solution no longer produces a distinction between days that have sunny mornings-cloudy afternoons and cloudy mornings-sunny afternoons. This is clearly an undesirable result since there are in fact days in Durban that have diurnal patterns of Class C and D, and in order to forecast, the classes that best represents the irradiance patterns in Durban should be used. Furthermore, even though the diurnal patterns are clearly distinguishable in the $\bar{D}_n$ mean profiles, it is beam irradiance that is most closely related to cloud cover and hence the best description of their patterns is required.

From the clustering of several irradiance variables it was found that they produce varying patterns for describing the irradiance in Durban. Some variables are able to produce well-defined clusters and hence distinct mean profiles, while others are not. In addition, their clustering can produce different solutions when clustered separately, as was seen with V_B, and in combinations, as seen with $\{\bar{B}_n, V_B\}$ and $\{\bar{B}_n, \bar{D}_n\}$. Having a multi-variable did not necessarily produce a better clustering solution, and in the case of $\{\bar{B}_n, \bar{D}_n\}$, it results in less distinguishable profiles. From all the variables clustered, $\bar{B}_n$ (derived from B_n), on its own has shown to be the best clustering variable in for classifying and characterizing irradiance patterns and therefore to be used for forecasting.

7.2 Forecasting using classes

7.2.1 Day forecasts of $\bar{B}_n$

As mentioned previously, a novelty of this study was clustering of Q profiles into classes in a similar manner to the radiometric variables. As explained in Chapter 6, in order to use classes of Q to forecast the irradiance class, it was necessary to have four classes of cloud cover. The class mean profiles of Q in Figure 6.2 also show a correspondence with the $\bar{B}_n$ profiles. More specifically, Q is low for Class A′ which corresponds to the $\bar{B}_n$ Class A of sunny days. Q is high for Class B′ corresponding to $\bar{B}_n$ Class B of cloudy days. Class C′ corresponds to the sunny AM-cloudy PM $\bar{B}_n$ Class C. Lastly, although Class D′ has a less prominent diurnal trend, the days contained therein still correspond to the cloudy AM-sunny PM $\bar{B}_n$ Class D. These were used by both forecasting methods.

As explained in Chapter 5, forecasting the $\bar{B}_n$ class for the day ahead was done using two methods, although both methods used the Q forecast from AccuWeather. The forecasting results using the classes of Q presented in Table 6.2, show that overall the method has a moderate success rate. Table 6.2 shows that prediction success rate per class ranges from 50%-72%, with Classes A and B having the highest rates. Regarding incorrect predictions, predicted Class A (sunny all day) was actually Class B (cloudy all day) for only one day (3%) and predicted Class B was actually Class A in zero cases. This makes sense because for a good NWP it should be unusual for a sunny day to be wrongly forecast as a cloudy day and vice-versa. Classes C and D had a higher percentage of incorrect predictions, although Class D's success rate of 58% is not much lower than that of Class B. The success rate of 65% over all classes may only be applied to the period January to June from which the sample of 100 days was drawn, but it so happens that a standardized rate based on weighting with class frequencies over one year, given in Table 5.2, also gives a success rate of 65%.

The second forecasting method uses the 50% rule as described in Table 6.3 in Chapter 6 and the results produced in Table 6.4 also show a fairly moderate overall success rate. In fact, the success rate of this method differs only by 2% from the success rate of the first method. Table 6.4 shows that the 50% rule produces prediction success rate per class in the range 59%-83%, which is somewhat better than that of the Q clustering forecast, except that there is a higher rate (9%) of Class B incorrectly predicted as Class A. The Class D success rate was significantly better, although the sample is rather small. Furthermore, Class D predicted by the 50%rule has only about a third of the days predicted to be in Class D by Q clustering because, as may be seen in Figure 6.1, Class D′ has a

large number of days with $C_{AM} \leq 50\%$ which are therefore predicted by the 50% rule to be in Class A. This increased the number of sunny (Class A) days predicted by the 50% rule. A consequence is that the 50% rule results in more Class A days being incorrectly predicted than in the Q clustering method. In comparison with Q clustering, the prediction success rate decreased for Classes A and B (by 8% and 7%, respectively) and increased for Classes C and D (by 13% and 25%, respectively). The overall raw success rate is 63%, and the standardized success rate is 64%, both of which, apart from being almost identical, are very close to that of the Q clustering method.

Table 6.5 shows the average RMSE values for the two day-ahead forecasting methods. Q clustering showed the best performance in predicting sunny days (Class A), followed by cloudy days (Class B). This may be due to the NWP model being able to distinguish between cloud-free and cloudy situations relatively well. By contrast, the 50% rule had a success rate that is better for the mixed conditions of Classes C and D. This may be due to the stronger separation of cloud cover conditions by the 50% rule as compared with clustering. Average profile error as quantified by RMSE was in the range 0.2-0.34. Overall, the average RMSE per class are fairly similar using the two forecasting methods.

These results are similar to those of Badosa et al. (2015), where the lowest RMSE was for sunny conditions. However, in contrast to the present study where the highest RMSE was found for mixed conditions, Badosa et al. (2015) found the highest forecasting errors were associated with cloudy conditions. The main difference in the methods is that the present work used cloud cover output of NWP, with clustering of cloud cover profiles, and uses beam rather than global irradiance.

The forecasting methods presented in this work have moderate success, which may be attributed both to the degree of accuracy of NWP and the existence of clusters of diurnal irradiance profiles which show a high degree of clustering for sunny and cloudy conditions but are less well-clustered for mixed conditions.

The novelty of this investigation was the use of clustering of cloud cover output from NWP, as well as the 50% rule, for day-ahead forecasts of beam irradiance using classification of daily irradiance profiles. Although previous studies such as Jeanty et al. (2013), Badosa et al. (2013) and Zhandire (2017) used clustering of irradiance to produce classes, they were not combined with cloud cover. The study by Zagouras et al. (2013) applied clustering to satellite-derived cloud estimates, but in contrast with the present work, this was for several geographical locations. McCandless et al. (2015) considered several outputs from the NWP including cloud cover, but were not used for

forecasting.

7.2.2 Hourly forecasts of $\bar{B}_n$ and $\bar{D}_n$

Intra-day forecasts of $\bar{B}_n$ and $\bar{D}_n$ were done at hourly intervals using the PCT method. The RMSE for $\bar{B}_n$ and $\bar{D}_n$ in each hour in all classes is given in Figure 6.9 (a) and (b), respectively. Class D showed the highest RMSE in $\bar{B}_n$ for most hours of the day, but are lower than other classes for $\bar{D}_n$. Classes A, B and C have RMSE of at most 0.25 for all hours. In most cases, for $\bar{D}_n$, Classes B, C and D show the highest RMSE. The variance of the classes in Figure 6.10 are the highest for A and B in $\bar{B}_n$ and $\bar{D}_n$.

The PCT method was compared to traditional Persistence and in some cases was found to show an improvement. Using the PCT method, the average improvement in RMSE over Persistence in $\bar{B}_n$ for all classes range from 6%-43% and in $\bar{D}_n$ from 4%-31%. For all classes, the average RMSE improvement of the PCT method over Persistence was found to be 22% for $\bar{B}_n$ and $\bar{D}_n$. This can be compared with an average percentage improvement of 18% achieved by McCandless et al. (2015) which also used cloud regime classes for forecasting.

7.2.3 General summary of classification and forecasting results

From the classification and clustering results using the *k*-means clustering method, beam irradiance was found to be the most appropriate variable for clustering since it was able to distinguish between sky conditions sufficiently well. Diffuse irradiance is another potentially useful quantity for describing sky conditions and a more in depth investigation of their combination with beam irradiance, could be considered for future studies. Furthermore, sub-hourly variability in the diffuse irradiance could be investigated as another possibility.

From the forecasting results of beam and diffuse irradiance, both methods using the cloud cover from the NWP were shown to have moderate success. To improve the performance of the forecasting methods, an opportunity for future work may include the use of cloud imagery.

Chapter 8

Conclusion

Solar power plant operators have the problem of dealing with the variable nature of solar irradiance, which impacts grid stability and reliability and affects activities such as load following and management, unit commitment and maintenance scheduling. This emphasizes the need for forecasting the amount of irradiance that would be available to the power plant at a certain time and therefore minimizing and possibly eliminating disturbances in the power output.

One of the essential steps to developing a forecasting model is to first have a proper understanding of the solar irradiance patterns at the given location. This study used the approach of clustering to understand, classify and characterize irradiance patterns. Clustering was applied to normalized hourly beam irradiance profiles ($\bar{B}_n$) in Durban, South Africa between 8:30 and 16:30 for 365 days during January 2014-January 2015. Results from the clustering yielded four $\bar{B}_n$ classes with distinct diurnal mean profiles that characterize the irradiance patterns for Durban. These were Class A: sunny all day, Class B: cloudy all day, Class C: sunny morning-cloudy afternoon and Class D: cloudy morning-sunny afternoon. In addition, there was a set of associated normalized diffuse irradiance profiles ($\bar{D}_n$) that describe the diurnal diffuse patterns.

The $\bar{B}_n$ irradiance classes were associated with predicted cloud cover percentage (Q) from the Numerical Weather Prediction for day-ahead forecasts. Clustering of Q was performed to obtain four classes with diurnal patterns associated with the $\bar{B}_n$ classes. Two forecasting methods were applied to forecast the class of $\bar{B}_n$ and $\bar{D}_n$. The first used the Q classes to forecast associated $\bar{B}_n$ classes, and the second used the 50% rule. The forecasting results showed that the two methods produced comparable prediction success rates in the range 50%-83%, with overall success rate about 65% for both methods. The Q clustering method showed the best performance in predicting sunny

days, followed by cloudy days. On the other hand, the 50% rule had a success rate that was better for the mixed cloud conditions of Classes C and D. Average profile error as quantified by Root Mean Square Error (RMSE) was in the range 0.2-0.34.

Hourly forecasts of $\bar{B}_n$ and $\bar{D}_n$ for the day ahead were produced using the Persistence of the Class Trend (PCT) method. The PCT method also used the Q forecast and the 50% rule to forecast an irradiance class. Thereafter, hour-ahead forecasts of $\bar{B}_n$ and $\bar{D}_n$ were performed using the class mean profiles to extrapolate to the next hour using the measured value at the current hour. Overall, for all classes, the PCT method showed an improvement over Persistence of approximately 22% in $\bar{B}_n$ and $\bar{D}_n$.

The clustering results presented in this work provide a classification of beam irradiance profiles for Durban, and a novel approach to day-ahead forecasting using classification of cloud cover predictions. Day-ahead forecasts have value in predicting the general daily profile, and are potentially useful for constraining models for multi-hour predictions on a particular day.

Bibliography

Abdulmouti, H. and Mansour, T. M. (2006). '**The technique of PIV and its applications**'. Proceedings of the 10th International Conference on Liquid Atomization and Spray Systems (ICLASS), Kyoto, Japan pp. 1–10.

Aguiar, L., Pereira, B., Lauret, P., Díaz, F. and David, M. (2016). '**Combining solar irradiance measurements, satellite-derived data and a numerical weather prediction model to improve intra-day solar forecasting**'. *Renewable Energy* **97**, 599–610.

Atwater, M. A. and Ball, J. T. (1981). '**Effects of clouds on insolation models**'. *Solar Energy* **27**, 37–44.

Badosa, J., Haeffelin, M. and Chepfer, H. (2013). '**Scales of spatial and temporal variation of solar irradiance on Reunion tropical island**'. *Solar Energy* **88**, 42–56.

Badosa, J., Haeffelin, M., Kalecinski, N., Bonnardot, F. and Jumaux, G. (2015). '**Reliability of day-ahead solar irradiance forecasts on Reunion Island depending on synoptic wind and humidity conditions**'. *Solar Energy* **115**, 306–321.

Benmouiza, K. and Cheknane, A. (2013). '**Forecasting hourly solar radiation using hybrid k-means and non linear autoregressive neural network models**'. *Energy Conversion and Management* **75**, 561–569.

Bramer, M. (2007). *Principles of data mining*. Springer, London.

Brooks, M., du Clou, S., van Niekerk, W., Gauché, P., Leonard, C., Mouzouris, M., Meyer, R., van der Westhuizen, N., van Dyk, E. and Vorster, F. (2015). '**SAURAN: A new resource for solar radiometric data in Southern Africa**'. *Journal of Energy in Southern Africa* **26 (1)**, 2–10.

Calbó, J., González, J. and Pagés, D. (2001). '**A method for sky-condition classification from ground-based solar radiation measurements**'. *Applied Meteorology* **40**, 2193–2199.

Calbó, J. and Sabburg, J. (2008). '**Feature extraction from whole-sky ground-based images for cloud-type recognition**'. *Atmospheric and Oceanic Technology* **25**, 3–14.

Cao, J. and Cao, S. (2006). '**Study of forecasting solar irradiance using neural networks with preprocessing sample data by wavelet analysis**'. *Energy* **31**, 3435–3445.

Cao, J. and Lin, X. (2008). '**Application of the diagonal recurrent wavelet neural network to solar irradiation forecast assisted with fuzzy technique**'. *Engineering Applications of Artificial Intelligence* **21**, 1255–1263.

Cao, S. and Cao, J. (2005). '**Forecast of solar irradiance using recurrent neural networks combined with wavelet analysis**'. *Applied Thermal Engineering* **25**, 161–172.

Castro-Almazán, J., Varela, A. and Munõz-Tuñóna, C. (2015). '**Day time cloud cover at Teide Observatory**'. Canarian Observatories Updates, Institute of Astrophysics of the Canary Islands (IAC) , pp. 1–4.

Cazorla, A., Olmo, F. and Alados-Arboledas, L. (2008). '**Development of a sky imager for cloud cover assessment**'. *Optical Society of America A* **25**, 29–39.

Chaâbane, M., Masmoudi, M. and Medhioub, K. (2004). '**Determination of Linke turbidity factor from solar radiation measurement in northern Tunisia**'. *Renewable Energy* **29**, 20652076.

Chaturvedi, D. (2016). '**Solar power forecasting: a review**'. *International Journal of Computer Applications* **145 (6)**, 28–50.

Chow, C., Urquhart, B., Lave, M., Dominguez, A., Kleissl, J., Shields, J. and Washom, B. (2011). '**Intra-hour forecasting with a total sky imager at the UC San Diego solar energy testbed**'. *Solar Energy* **85**, 2881–2893.

Crispim, E., Ferreira, P. and Ruano, A. (2008). '**Prediction of the solar radiation evolution using computational intelligence techniques and cloudiness indices**'. *Innovative Computing Information and Control* **4**, 1121–1133.

Cros, S., Sébastien, N., Liandrat, O., Jolivet, S. and Schmutz, N. (2014). '**Solar power forecasting for a massive and secure injection of photovoltaics on the grid**'. Proceedings of the 4th International Workshop on Integration of Solar Power into Power Systems, Berlin, Germany, pp. 2–4.

Dambreville, R., Blanc, P., Chanussot, J., Boldo, D. and Dubost, S. (2014). '**Very short term forecasting of the global horizontal irradiance through Helioclim maps**'. Proceedings of the 5th International Renewable Energy Congress (IREC), Hammamet, Tunisia, pp. 1–6.

Davies, J. A. and McKay, D. C. (1982). '**Estimating solar irradiance and components**'. *Solar Energy* **29**, 55–64.

Dedekind, Z., Engelbrecht, F. A. and van der Merwe, J. (2016). '**Model simulations of rainfall over southern Africa and its eastern escarpment**'. *Water SA* **42 (1)**, 129–143.

Diabaté, L., Blanc, P. and Wald, L. (2004). '**Solar radiation climate in Africa**'. *Solar Energy* **76**, 733–744.

Diagne, M., David, M., Lauret, P., Boland, J. and Schmutz, N. (2013). '**Review of solar irradiance forecasting methods and a proposition for small-scale insular grids**'. *Renewable and Sustainable Energy Reviews* **27**, 65–76.

Duffie, J. and Beckman, W. (1991). *Solar engineering of thermal processes*. John Wiley and Sons, New York.

Eastman, R. and Warren, S. G. (2013). '**Diurnal cycles of cumulus, cumulonimbus, stratus, stratocumulus, and fog from surface observations over land and ocean**'. *Journal of Climate* **27**, 2386–2404.

Ferreira, P., Gomes, J., Martins, I. and Ruano, A. (2012). '**A neural network based intelligent predictive sensor for cloudiness, solar radiation and air temperature**'. *Sensors* **12**, 15750–15777.

Feussner, K. and Dubois, P. (1930). '**Trübungsfactor, precipitable water Staub.**'. *Gerlands Beitr. Geophys.* **27**, 132–175.

Gastón-Romeo, M., Leon, T., Mallor, F. and Ramírez-Santigosa, L. (2011). '**Morphological clustering method for daily solar radiation curves**'. *Solar Energy* **85**, 1824–1836.

Goswami, D., Kreith, F. and Kreider, J. (1999). *Principles of solar engineering*. Taylor and Francis, Philadelphia.

Grenier, J. C., Casiniere, A. D. L. and Cabot, T. (1995). '**Atmospheric turbidity analyzed by means of standardized Linke's turbidity factor**'. *Journal of Applied Meteorology* **34**, 14491458.

Gueymard, C. A. (1989). '**A two-band model for the calculation of clear sky solar irradiance, illuminance, and photosynthetically active radiation at the earth's surface**'. *Solar Energy* **43**, 253–265.

Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001). '**On clustering validation techniques**'. *Intelligent Information Systems* **17**, 107–145.

Hammer, A., Heinemann, D., Lorenz, E. and Lückehe, B. (1999). '**Short-term forecasting of solar radiation: a statistical approach using satellite data**'. *Solar Energy* **67**, 139–150.

Harrouni, S., Guessoum, A. and Maafi, A. (2005). '**Classification of daily solar irradiation by fractional analysis of 10-min-means of solar irradiance**'. *Theoretical and Applied Climatology* **80**, 27–36.

Haurwitz, B. (1948). '**Insolation in relation to cloud type**'. *Journal of Meteorology* **3**, 110113.

Huo, J. and Lu, D. (2009). '**Cloud determination of all-sky images under low-visibility conditions**'. *Atmospheric and Oceanic Technology* **26**, 2172–2181.

Ineichen, P. (2006). '**Comparison of eight clear sky broadband models against 16 independent data banks**'. *Solar Energy* **80**, 468–478.

Ineichen, P. and Perez, R. (2002). '**A new airmass independent formulation for the Linke turbidity coefficient**'. *Solar Energy* **73**, 151–157.

Inman, R., Pedro, H. T. and Coimbra, C. (2013). '**Solar forecasting methods for renewable energy integration**'. *Progress in Energy and Combustion Science* **39**, 535–576.

Jacovides, C. P. (1997). '**Model comparison for the calculation of Linke's turbidity factor**'. *International Journal of Climatology* **17**, 551563.

Jeanty, P., Delsaut, M., Trovalet, L., Ralambondrainy, H., Lan-Sun-Luk, J., Bessafi, M., Charton, P. and Chabriat, J. (2013). '**Clustering daily solar radiation from Reunion Island using data analysis methods**'. Proceedings of the International Conference on Renewable Energies and Power Quality (ICREPQ'13), Bilbao, Spain .

Johnson, R., Hering, W. and Shields, J. (1989). '**Automated visibility and cloud cover measurements with a solid-state imaging system**'. Scripps Institution of Oceanography, Marine Physical Laboratory, University of California, San Diego .

Jolliffe, I. (2002). *Principal Component Analysis*. Springer, New York.

Kalogirou, S. (2009). *Solar energy engineering: processes and systems*. Academic Press, USA.

Kamath, C. (2010). '**Understanding wind ramp events through analysis of historical data**'. Proceedings of IEEE PES Transmission and Distribution Conference, Los Angeles, pp. 3–8.

Kang, B. and Tam, K. (2013). '**A new characterization and classification method for daily sky conditions based on ground-based solar irradiance measurement data**'. *Solar Energy* **94**, 102–118.

Kassianov, E., Long, C. and Ovtchinnikov, M. (2005). '**Cloud sky cover versus cloud fraction: whole-sky simulations and observations**'. *Applied Meteorology* **44**, 86–98.

Kasten, F. (1980). '**A simple parameterization of two pyrheliometric formulae for determining the Linke turbidity factor**'. *Meteor. Rdsch* **33**, 124–127.

Kasten, F. and Young, A. (1989). '**Revised optical air mass tables and approximation formula**'. *Applied Optics* **28**, 4735–4738.

Kaufman, L. and Rousseeuw, P. (1990). *Finding groups in data: an introduction to cluster analysis*. Wiley, New York.

Kazantzidis, A., Tzoumanikas, P., Bais, A., Fotopoulos, S. and Economou, G. (2012). '**Cloud detection and classification with the use of whole-sky ground-based images**'. *Atmospheric Research* **113**, 80–88.

Kipp and Zonen (2014). '**CHP1 pyrheliometer instruction manual**'.
URL: *http://www.kippzonen.com/Product/18/CHP1-Pyrheliometer.* Accessed June 2016.

Kleissl, J. (2013). *Solar energy forecasting and resource assessment*. Academic Press, Massachusetts.

Kostylev, V. and Pavlovski, A. (2012). '**Solar power forecasting performance-towards industry standards**'. First International Workshop on the Integration of Solar Power into Power Systems, Denmark .

Lam, J. C. and Li, D. H. W. (1998). '**Correlation analysis of solar radiation and cloud cover**'. *International Journal of Ambient Energy* **19 (4)**, 187–198.

Lauret, P., Lorenz, E. and David, M. (2016). '**Solar forecasting in a challenging insular context**'. *Atmosphere* **7**, 117.

Linke, F. (1922). '**Transmissions-Koeffizient und Trübungsfaktor**'. *Beitr. Phys. fr. Atmos.* **10**, 91–103.

Lletí, R., Ortiz, M., Sarabia, L. and Sánchez, M. (2004). '**Selecting variables for *k*-means cluster analysis by using a genetic algorithm that optimises the silhouettes**'. *Analytica Chimica Acta* **515**, 87–100.

Lohmann, U., Lüönd, F. and Mahrt, F. (2016). *An introduction to clouds: from the microscale to climate*. Cambridge University Press, UK.

Long, C., Calbó, J., Sabburg, J. and Pagés, D. (2006). '**Retrieving cloud characteristics from ground-based daytime color all-sky images**'. *Atmospheric and Oceanic Technology* **23**, 633–652.

Lorenz, E., Hammer, A. and Heinemann, D. (2004). '**Short term forecasting of solar radiation based on satellite data**'. Proceedings of ISES Europe Solar Congress, Freiburg, Germany, pp. 1–8.

Louche, A., Peri, G. and Iqbal, M. (1986). '**An analysis of Linke turbidity factor**'. *Solar Energy* **37**, 393396.

Lysko, M. (2006). *Measurement and models of solar irradiance*. PhD Thesis, Norwegian University of Science and Technology, Faculty of Natural Sciences and Technology.

Maafi, A. and Harrouni, S. (2003). '**Preliminary results of the fractal classification of daily solar irradiances**'. *Solar Energy* **75**, 53–61.

MacQueen, J. (1967). '**Some methods for classification and analysis of multivariate observations**'. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Volume I: Statistics pp. 281–297.

Marquez, R. and Coimbra, C. (2013). '**Intra-hour DNI forecasting methodology based on cloud tracking image analysis**'. *Solar Energy* **91**, 327–336.

Martin, L., Zarzalejo, L., Polo, J., Navarro, A., Marchante, R. and Cony, M. (2010). '**Prediction of global solar irradiance based on time series analysis: application to solar thermal power plants energy production planning**'. *Solar Energy* **84**, 1772–1781.

Martínez-Chico, M., Batlles, F. and Bosch, J. (2011). '**Cloud classification in a Mediterranean location using radiation data and sky images**'. *Energy* **36**, 4055–4062.

Mathiesen, P. (2013). *Advanced numerical weather prediction techniques for solar irradiance forecasting: statistical, data-assimilation, and ensemble forecasting*. PhD Thesis, University of California, San Diego.

Mathiesen, P. and Kleissl, J. (2011). '**Evaluation of numerical weather prediction for intra-day solar forecasting in the continental United States**'. *Solar Energy* **85**, 967–977.

McCandless, T., Haupt, S. and Young, G. (2014). '**Short term solar radiation forecasts using weather regime-dependent artificial intelligence techniques**'. Proceedings of the 12th Conference on Artificial and Computational Intelligence and its Applications to the Environmental Sciences: Applications of Artificial Intelligence Methods for Energy, Atlanta, Georgia .

McCandless, T., Haupt, S., Young, G. and Annunzio, A. (2015). '**A regime dependent bayesian approach to short-term solar irradiance forecasting**'. Proceedings of the 13th Conference on Artificial Intelligence: Applications of Artificial Intelligence Methods for Energy-Part II, Phoenix, Arizona .

Mellit, A. (2008). '**Artificial Intelligence technique for modelling and forecasting of solar radiation data: a review**'. *Artificial Intelligence and Soft Computing* **1**, 52–76.

Miller, D. H. (1981). *Energy at the surface of the Earth*. Academic Press, New York.

Muller, M. (1995). '**Equation of time-problem in astronomy**'. *Acta Physica Polonica A* **88**, 1–18.

Muneer, T. (1997). *Solar radiation and daylight models*. Elsevier, USA.

Muselli, M., Poggi, P., Notton, G. and Louche, A. (2000). '**Classification of typical meteorological days from global irradiation records and comparison between two Mediterranean coastal sites in Corsica Island**'. *Energy Conversion and Management* **41**, 1043–1063.

Muselli, M., Poggi, P., Notton, G. and Louche, A. (2001). '**First order Markov chain model for generating synthetic "typical days" series of global irradiation in order to design photovoltaic stand alone systems**'. *Energy Conversion and Management* **42**, 675–687.

Myers, D. (2005). '**Solar radiation modeling and measurements for renewable energy applications: data and model quality**'. *Energy* **30**, 1517–1531.

Myers, D. (2013). *Solar radiation: practical modeling for renewable energy applications*. Taylor and Francis, Florida.

Myers, D. R. and Wilcox, S. M. (2009). '**Relative accuracy of 1-minute and daily total solar radiation data for 12 global and 4 direct beam solar radiometers**'. American Solar Energy Society Annual Conference, Buffalo, New York, pp. 348–359.

Nel, W. (2009). '**Rainfall trends in the KwaZulu-Natal Drakensberg region of South Africa during the twentieth century**'. *International Journal of Climatology* **29**, 16341641.

Paoli, C., Voyant, C., Muselli, M. and Nivet, M. (2010). '**Forecasting of preprocessed daily solar radiation time series using neural networks**'. *Solar Energy* **84**, 2146–2160.

Paoli, C., Voyant, C., Muselli, M. and Nivet, M. (2014). '**Multi-horizon irradiation forecasting for Mediterranean locations using time series models**'. *Energy Procedia* **57**, 1354 – 1363.

Paulescu, M., Paulescu, E., Gravila, P. and Badescu, V. (2013). *Weather modeling and forecasting of PV systems operation*. Springer, London.

Pelland, S., Remund, J., Kleissl, J., Oozeki, T. and Brabandere, K. D. (2013). '**Photovoltaic and solar forecasting: state of the art**'. International Energy Agency (IEA) Report, pp. 1–36.

Perez, R., Kivalov, S., Schlemmer, J., Jr., K. H., Renné, D. and Hoff, T. (2010). '**Validation of short and medium term operational solar radiation forecasts in the US**'. *Solar Energy* **84**, 2161–2172.

Perez, R., Moore, K., Wilcox, S., Renne, D. and Zelenka, A. (2007). '**Forecasting solar radiation - Preliminary evaluation of an approach based upon the national forecast database**'. *Solar Energy* **81**, 809–812.

Pfister, G., McKenzie, R., Liley, J. and Thomas, A. (2003). '**Cloud coverage based on all-sky imaging and its impact on surface solar irradiance**'. *Applied Meteorology* **42**, 1421–1434.

Quesada-Ruiz, S., Chu, Y., Tovar-Pescador, J., Pedro, H. and Coimbra, C. (2014). '**Cloud-tracking methodology for intra-hour DNI forecasting**'. *Solar Energy* **102**, 267–275.

Réchou, A., Rao, T. N., Bousquet, O., Plu, M. and Decoupes, R. (2014). '**Properties of rainfall in a tropical volcanic island deduced from UHF wind profiler measurements**'. *Atmospheric Measurement Techniques* **7**, 409418.

Reikard, G. (2009). '**Predicting solar radiation at high resolutions: a comparison of time series forecasts**'. *Solar Energy* **83**, 342–349.

Remund, J., Lefevre, M., Ranchin, T. and Page, J. (2003). '**Worldwide Linke turbidity information**'. Proceedings of ISES Solar World Congress, Göteborg, Sweden p. 13.

Remund, J., Perez, R. and Lorenz, E. (2008). '**Comparison of solar radiation forecasts for the USA**'. Proceedings of the 23rd European Photovoltaic Solar Energy Conference, Valencia, Spain, pp. 1–3.

Reno, M. (2012). '**Pvlib toolbox for Matlab. Sandia National Laboratories**'.
URL: *http://Pvlib toolbox matlab.pvpmc.sandia.gov/applications/pvlib-toolbox/matlab/.*
Accessed August 2017

Reno, M., Hansen, C. and Stein, J. (2012). '**Global horizontal irradiance clear sky models: implementation and analysis**'. Sandia National Laboratories. Sandia Report (SAND2012-2389), pp. 1–68.

Reno, M. J. and Hansen, C. W. (2016). '**Identification of periods of clear sky irradiance in time series of GHI measurements**'. *Renewable Energy* **90**, 520–531.

Rogers, R. R. and Yau, M. K. (1989). *A short course in cloud physics*. Elsevier, USA.

Rousseeuw, P. (1957). '**Silhouettes: a graphical aid to the interpretation and validation of cluster analysis**'. *Computational and Applied Mathematics* **20**, 53–65.

SAURAN (2014). '**Southern African Universities Radiometric Network**'.
URL: *http://www.SAURAN.net*

Sen, Z. (2008). *Solar energy fundamentals and modeling techniques*. Springer, London.

Shields, J., Johnson, R. and Koehler, T. (1993). '**Automated whole sky imaging systems for cloud field assessment**'. Fourth Symposium on Global Change Studies, Anaheim, Canada, pp. 228–231.

Shields, J., Karr, M. and R.W. Johnson, A. B. (2013). '**Day/night whole sky imagers for 24-h cloud and sky assessment: history and overview**'. *Applied Optics* **52**, 1605–1616.

SoDa (2011). '**Solar Radiation Data Service**'.
URL: *http://www.soda-is.com/eng/index.html.* Accessed August 2017.

Soubdhan, T., Emilion, R. and Calif, R. (2009). '**Classification of daily solar radiation distributions using a mixture of Dirichlet distributions**'. *Solar Energy* **83**, 1056–1063.

South African Weather Service (2010). '**Aeronautical climatological summary (1996-2010)**'.
URL: *http://www.weathersa.co.za.* Accessed September 2014.

Tapakis, R. and Charalambides, A. (2013). '**Equipment and methodologies for cloud detection and classification: a review**'. *Solar Energy* **95**, 392–430.

Tufféry, S. (2011). *Data mining and statistics for decision making*. John Wiley and Sons, UK.

Twidell, J. and Weir, T. (2006). *Renewable energy resources*. Taylor and Francis, New York.

Voyant, C., Muselli, M., Paoli, C. and Nivet, M. (2011). '**Optimization of an artificial neural network dedicated to the multivariate forecasting of daily global radiation**'. *Energy* **36 (1)**, 348–359.

Voyant, C., Muselli, M., Paoli, C. and Nivet, M. (2012). '**Numerical weather prediction (NWP) and hybrid ARMA/ANN model to predict global radiation**'. *Energy* **39**, 341–355.

Voyant, C., Notton, G., Kalogirou, S., Nivet, M., Paoli, C., Motte, F. and Fouilloy, A. (2017). '**Machine learning methods for solar radiation forecasting: a review**'. *Renewable Energy* **105**, 569–582.

Ward, J. H. (1963). '**Hierarchical grouping to optimize an objective function**'. *Journal of the American Statistical Association* **58**, 236–244.

Watanabe, T., Takamatsu, T. and Nakajima, T. Y. (2016). '**Evaluation of variation in surface solar irradiance and clustering of observation stations in Japan**'. *Applied Meteorology and Climatology* **55**, 2165–2180.

WMO (2008). '**Guide to meteorological instruments and methods of observation**'. **URL:** *http://www.wmo.int/pages/prog/www/IMOP/CIMO-Guide.html.* Accessed October 2017.

Wood, R. (2012). '**Review: Stratocumulus Clouds**'. *Montly Weather Review* **140**, 23732423.

Zagouras, A., Inman, R. H. and Coimbra, C. (2014). '**On the determination of coherent solar microclimates for utility planning and operations**'. *Solar Energy* **102**, 173–188.

Zagouras, A., Kazantzidis, A., Nikitidou, E. and Argiriou, A. (2013). '**Determination of measuring sites for solar irradiance, based on cluster analysis of satellite-derived cloud estimations**'. *Solar Energy* **97**, 1–11.

Zangvil, A. and Lamb, P. (1997). '**Characterization of sky conditions by the use of solar radiation data**'. *Solar Energy* **61**, 17–122.

Zawilska, E. and Brooks, M. (2011). '**An assessment of the solar resource for Durban, South Africa**'. *Renewable Energy* **36**, 3433–3438.

Zhandire, E. (2015). 'Private communication'.

Zhandire, E. (2017). '**Solar resource classification in South Africa using a new index**'. *Journal of Energy in Southern* **28 (2)**, 61–70.