Spatial Statistical Modeling and Characterization of Aerosol Optical Thickness over Lahore using MODIS Data

A Master's Thesis
Presented by

Safa Ashraf

In Partial Fulfillment
of the Requirements of the Degree of
Master's in
Electrical Engineering

Supervisors:
Dr. Zubair Khalid
Dr. Momin Uppal
Dr. Muhammad Tahir

Department of Electrical Engineering
Syed Babar Ali School of Science and Engineering
Lahore University of Management Sciences
Spring 2019

This Master's Thesis has been examined by a Committee of the Department of Electrical Engineering as follows:

Dr. Abubakr Muhammad .....
Evaluator, Thesis Committee
Associate Professor, Electrical Engineering Department, LUMS

Dr. Zubair Khalid .....
Thesis Supervisor
Assistant Professor, Electrical Engineering Department, LUMS

Dr. Momin Uppal .....
Thesis Supervisor
Associate Professor, Electrical Engineering Department, LUMS

Dr. Muhammad Tahir .....
Thesis Supervisor
Assistant Professor, Electrical Engineering Department, LUMS

To my parents.

Acknowledgments

First off, I would like to acknowledge the constant support of my parents towards my academic ambitions throughout my life that has been a great motivation for me to reach this stage. I would further like to acknowledge the encouraging advice of my supervisor, Dr. Muhammad Tahir to consider getting enrolled in the graduate program at LUMS, which indeed proved to be a great learning experience. Let me appreciate the unwavering support of my supervisors, Dr. Momin Uppal, Dr. Muhammad Tahir and Dr. Zubair Khalid for their guidance, encouragement and patience throughout the journey of this research work; and their uplifting assurances for whenever I doubted my work. I would like to mention my favorite courses at LUMS: Stochastic Systems, Advanced Digital Signal Processing, and Information Theory and Machine Learning that broadened the horizons of my mind greatly; and my professors, Dr. Momin Uppal, Dr. Muhammad Tahir and Dr. Hassan Mohy Ud Din for always being there to assist my understanding of the course contents. Special thanks to Dr. Zubair Khalid for providing research grant from Higher Education Commission (HEC) under National Research Program for Universities (NRPU) Project Number 5925, 2016-17 to support this work and reviewing the first draft of this document thoroughly. Lastly, I acknowledge NASA's LAADS DAAC and AERONET for providing the data for free, without which, this work wouldn't have been possible.

Abstract

Increased concentration of aerosols in air caused by ever rising urbanization, development of various industries and natural phenomenon has horrendous consequences on human health, environment and climate. To counter these adverse effects of air pollution, spatial characterization of aerosol concentration in the region of interest needs to be developed first. For this purpose, Aerosol Optical Thickness (AOT) data product for the years 2017-2018, based on remote-sensing data of MODerate resolution Imaging Spectroradiometer (MODIS) on NASA's Aqua satellite was utilized having a spatial resolution of 3 KM and temporal resolution of 1-2 days. For the validity of this MODIS AOT data, positive correlation was established between AERONET AOT and MODIS AOT. Glowworm Swarm Optimization (GSO) algorithm was employed to identify several aerosol hot-spot locations in and around Lahore with their associated aerosol content based on MODIS AOT data. In addition to aerosol hot-spot characterization, a spatial statistical model based on Gaussian Processes with ARD Exponential kernel was also proposed to solve the problem of spatial coverage holes in MODIS data, using which, one can predict the value of AOT at any location in Lahore with a certain confidence level. Cross-validation results revealed the yearly normalized MSE of 0.0093 and 0.0106 for 2017 and 2018 respectively. These research directions were explored for the first time in the context of air pollution studies for Lahore and provide an interesting insight on regional aerosol concentration of Lahore for air quality management.

Contents

List of Figures

Chapter 1

Introduction

In recent times, an enormous trend towards urbanization and development of various industries in developing countries like Pakistan is evident. This ever rising urbanization has brought with itself the crisis of air pollution with absolutely horrendous consequences. The severity of this crisis is reflected in the 2015 World Health Organization (WHO) report, according to which, almost 60,000 people died from large amounts of fine particles in air and is one of the highest death tolls recorded globally due to air pollution (FAO of UN report, 2018). Moreover, according to a statement available from Punjab's Environment Protection Department (EPD), the average air pollution in Pakistan's big cities is up to four times higher than the WHO limits.

These aforementioned figures are more than enough of a reason to motivate policy decisions to counter the harmful effects of air pollution. However, in order to ensure that the right measures are put in place, the exact scale of the problem needs to be understood. Thus we need to develop a precise spatial and temporal characterization of air pollutants and identify the major sources from which they originate. Unfortunately, such a characterization in the context of Pakistani environment is non-existent. There are positive signs however. In the last few

years, the EPD has deployed air quality sensors at several locations in Lahore with those sensor readings made publicly available [1]. To monitor data from more locations, a couple of privately owned sensors have been recently deployed at various locations as well. In addition to these, AErosol RObotic NETwork (AERONET) sensors installed in Lahore and Karachi provide data of useful atmospheric aerosol properties like Aerosol Optical Thickness (AOT). However, these ground-based sensors provide measurements for selected cities such as Lahore, and that too with significant coverage holes. Even if the ground-based measurement nodes within Lahore were of satisfactory spatial density, determining an elaborate characterization of pollutants using only ground-based sensor measurements becomes infeasible as it requires an elaborate sensor network covering large swathes of geographical areas. Fortunately, satellite-based remote sensing of aerosols provides an attractive alternative that provide a global coverage. While there are several satellites that provide aerosol measurements, MODIS, a sophisticated instrument on NASA's Aqua and Terra satellites, with a promising spatial resolution of 10 km and 3 km happens to provide the most accurate and reliable satellite-data of aerosol properties for Lahore [2].

Using both the satellite and ground-based sensors, researchers have been trying to understand the air pollution sources and trends in Pakistan over the last two decades. Research in this domain has previously investigated the problem in the following directions: Seasonal trends in the properties of aerosols in Lahore including the variation in AOT values and classification of aerosols into two major types (desert dust and biomass burning/urban industrial) based on bi-modal distributions of Angstrom Exponent and AOT from Aqua MODIS and AERONET [3]. Spatio-temporal variations in aerosol concentration for several cities of Pakistan was analyzed using the Hybrid Single Particle Lagrangian Integrated Trajectory (HYSPLIT) model on satellite-based data of MODIS, Multi-angle Imaging Spec-

troradiometer (MISR) and Total Ozone Mapping Spectrometer (TOMS) via back-tracking air mass trajectories [4], [5], [6]. The above two research directions have revealed that the average AOT concentration in pre-monsoon (or summer) is the highest followed with that in the monsoon season, with the lowest value in winter. Moreover, classification analysis discovered the presence of coarse-mode particles with dominance in the pre-monsoon season owing to high dust storm frequency, while post-monsoon with fine particles, winter and monsoon with both fine and coarse-mode particles.

Other research directions have explored the validity of satellite-based AOT measurements with the ground-based AERONET data using correlation analysis [7]. These studies have compared the reliability of the two most important MODIS AOT retrieval algorithms, Deep Blue (DB) [8] and Dark Target (DT) [9] for Karachi and Lahore [10]. It was also discovered that DT and DB data products are more suitable for the urban areas in Pakistan dominated by coarse and fine particles respectively. However, AOT based on DT is the most suitable to use for Lahore as the region around Karachi has more bright surfaces and requires algorithms like DB to retrieve AOT on those regions [11].

In addition to the above literature, there are some studies that were conducted on the ground-based sensor measurements only. For example, there is an elaborate study of various aerosol properties like AOT, Angstrom Exponent (AE), Single Scattering Albedo (SSA), Aerosol Radiative Forcing (ARF) over Lahore using only the ground-based AERONET data [12]. Moreover, ground-based samplers have also been used to detect and quantify the presence of certain particles in the air. This approach was followed to carry out the detection of various atmospheric trace metals present in Islamabad's air, which showed the highest amount for Iron with $1.761\,\mu g\,m^{-3}$ followed by Sodium with $1.661\,\mu g\,m^{-3}$[13].

The variation in concentration of aerosol over a region tends to have an impact

on its climate. To understand this relationship, an analysis based on correlation between properties of aerosol and clouds was performed over 8 big cities of Pakistan using the aerosol and atmosphere data products of MODIS onboard Terra [14]. Water vapor, cloud fraction, cloud top temperature and cloud top pressure were the four important optical properties of clouds that were considered in the analysis. In another study, heavy pollution episodes were classified into dust episode (DE) and haze episode (HE) over Karachi and Lahore using correlation between AOD and AE [15].

After having gone through the relevant research work, it is not hard to observe that no significant contribution has been made to identify and quantify the air pollution sources in terms of aerosol hot-spots in Pakistan. This is indeed a very important step that needs to be investigated before making any effort to bring the air quality back into the safe limits. This is because only with the help of such aerosol hot-spots characterization, one would be able figure out the locations that need the most attention with regards to the air quality control. In addition to this, an interesting research direction of learning statistical model for aerosol properties can assist in better aerosol characterization and other relevant analyses by e.g. estimating the values for missing data. This kind of statistical modeling also seems non-existent in the context of Pakistan's air pollution research.

1.1 Outline and Contribution of the Thesis

1.1.1 Thesis Objectives

The objective of this thesis is to explore the following two research dimensions that were discerned above using data from Aqua's MODIS aerosol product.

Statistical Modeling of Aerosol Optical Thickness:

In this phase, a spatial statistical model, based on Gaussian Processes (GP) is proposed for Aerosol Optical Thickness (AOT), which is the most significant property of aerosols. The parameters for kernel governing the statistical relationship between the AOT at different locations can be learned based on likelihood maximization in GP framework. K-folds cross validation can be further employed to choose the most potential GP model. This model will provide a way to estimate the AOT value at any location in the region under consideration with a certain confidence level and thus, a better understanding of the spatio-temporal distribution of aerosol concentration can be inferred from it.

Aerosol Hot-spot Characterization:

In this part of the thesis, the regions in Lahore that have persistently high values of aerosols are identified using the Glowworm Swarm Optimization algorithm. After having identified the locations, the next step is to quantify the concentration of aerosols within these hot-spots using appropriate quantification metrics. At the end, hot-spots can be ranked according to their aerosol concentration. Since the locations are geographical coordinates, one can readily tell what areas of Lahore fall within the identified hot-spots with their estimated aerosol concentration.

1.1.2 Thesis Outline

In a nutshell, Chapter 2 presents the scientific background of atmospheric science, related preliminaries and problem formulation. Next two chapters provide a detailed theoretical description of the proposed methods to solve the problems identified at the end of Chapter 2. Chapter 3 is about the theoretical concepts which govern the working of Gaussian Processes Regression. While in Chapter

4, the mechanism of Glowworm Swarm Optimization algorithm is explained in a detail for the purpose of optimizing multi-modal functions. In order to achieve the objectives of the thesis, the concepts introduced in Chapter 3 and 4 are to be employed according to the nature of problem statements and MODIS data structure. The results of all the analyses performed on the data and findings after applying the proposed methods are presented in Chapter 5. Chapter 6 concludes the thesis by summarizing the results and highlighting future considerations.

Chapter 2

Preliminaries and Problem Formulation

This is an introductory chapter that discusses science of atmospheric pollution, types of pollutants and their hazards, causes that lead to high concentrations of pollutants and different techniques to measure these concentrations. The details of satellite data from MODIS that was used in this work is also described followed by the identification of research dimensions that were explored in this work.

2.1 Understanding Atmospheric Pollution

2.1.1 Scientific Background

The word atmosphere originated from two Greek words atmos meaning vapor and sphaira which means sphere. Atmosphere is a set of gaseous layers that surrounds the planet and is retained by its gravitational pull. The Earth's atmosphere has been divided into five layers depending on the composition, temperature and pressure of each layer. The layer that harbors life on Earth is called troposphere. This

layer is composed of different types of gases, suspended liquid droplets and various particulate matter. Among the gases, about 78% Nitrogen, 30% Oxygen, 0.93% Argon, 0.04% Carbon Dioxide and traces of other gases is present. This ideal composition is bound to change as a consequence of various phenomenon occurring on the planet's surface which become the sources of atmospheric pollution.

2.1.2 Sources of Atmospheric Pollution

Many natural phenomenon as well as anthropogenic activities are responsible for creating an unhealthy change in the composition of air. Natural sources of air pollution include volcanic eruption, acid rain, wild fires, sea-sprays, dust storms and biological allergens. However, in the context of developing countries such as Pakistan, degraded atmosphere is primarily due to anthropogenic sources such as vehicular fuel and gasoline combustion, industrial boilers, burning of coal and wood, commercial and residential heaters, waste disposal, chemically harmful sprays and paints. All these sources of air pollution can potentially increase the atmospheric concentration of Sulphur and Nitrogen oxides, hydrocarbons (e.g. Methane), Carbon Monoxide, ground-level Ozone, and even Lead.

Liquid droplets and small suspended particles present in the air form another significant class of pollutants called Particulate Matter (PM). These are of major interest in the context of this work. These are divided into different types based on their dimensions. Generally, they are classified in the two classes called PM_{10} and $PM_{2.5}$. PM_{10} include all those particles that are of dimensions less than or equal to 10 µm whereas $PM_{2.5}$ are of dimensions less than or equal to 2.5 µm. Figure 2-1 illustrates some well-known particles in PM against their dimensions in microns.

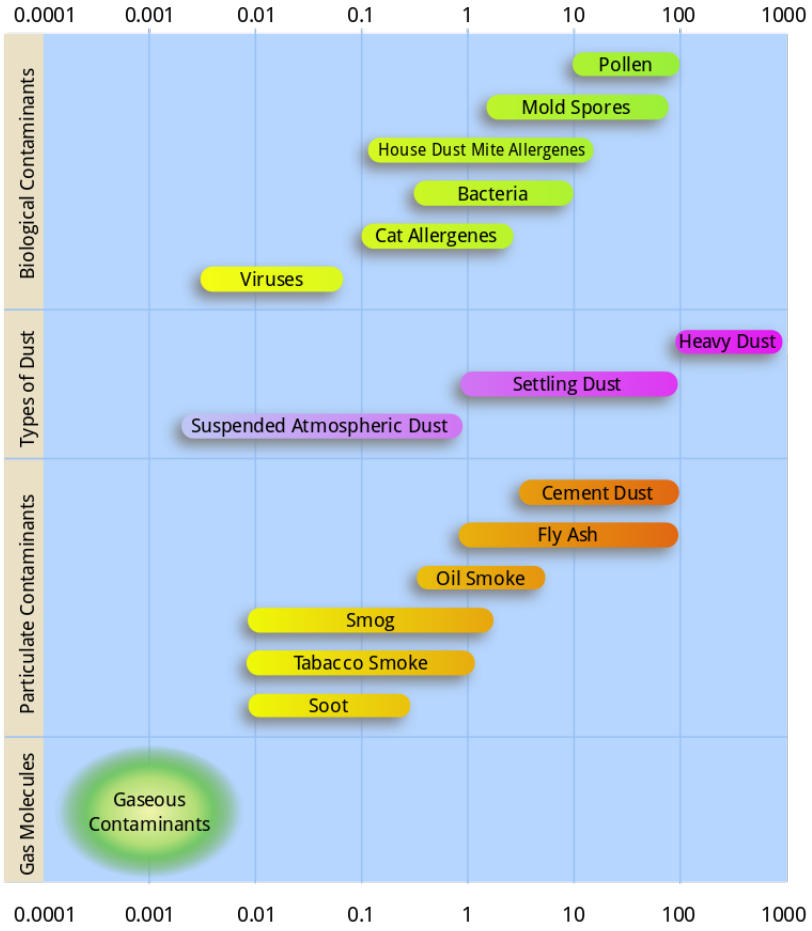

Figure 2-1: Particulate matter with their dimensions in microns (Image courtesy: https://en.wikipedia.org/wiki/Particulates).

2.1.3 Classification of Pollutants

Different pollutants that were mentioned earlier can be classified into two major categories based on how they enter into the atmosphere. They are usually divided into two major categories of primary and secondary pollutants. Primary pollutants are the ones that directly enter into the atmosphere and become a part of it. Carbon Monoxide coming out of the vehicles, Volatile Organic Compounds (VOCs) being ejected out of industries and all types of PM are some chief examples.

Secondary pollutants emerge when the primary pollutants combine together under some favorable conditions. Ozone is the most prominent example and is formed when Nitrogen Dioxide and VOCs react together. Sulphuric acid and Ammonia are some other examples of secondary pollutants.

Aerosol: The Pollutant under Consideration

Aerosol is suspension of solid particles and suspended liquid droplets in a gas, and in fact, just another term that is used to refer to the PM present in air. As mentioned earlier, these are the pollutants that are of major interest in this work. Their scientific properties can be exploited to quantify and even classify the suspended particles in the air. One of these properties, Aerosol Optical Thickness (AOT), happens to be the very fundamental measure based on which the entire analysis and modeling will be done. Description of the relevant properties are as follows:

- Aerosol Optical Thichness (AOT) is defined as a measure of quantity of light removed from the sunbeam reaching the surface of Earth. AOT inherently measures the quantity of particles present in the air that scatter the sunbeam by relating it to the intensity of light being observed at the surface. Thus, higher the AOT, higher will be amount of aerosols present in the air.

- Angstrom Exponent (AE) is defined as the negative slope of AOT with respect to wavelength. It is useful to infer about the sizes and hence, the class of aerosols.

2.1.4 Effects and Hazards of Pollutants

The aforementioned pollutants of atmosphere, even seeming unlikely, have horrendous consequences. These include contamination of crops and livestock, pro-

duction of greenhouse gases, acid rain, reduced visibility and formation of smog [16]. Moreover, pollutants can alter cloud properties by absorbing and reflecting sunlight, due to which the climate and the hydrological cycle is affected [17]. The biggest impact however is on the human body with cardiopulmonary health being the biggest sufferer. As far as PM is concerned, it is the major contributor to diseases of eyes and respiratory system. PM with smaller dimensions can even penetrate deep into the bloodstreams which can lead to many diseases [18]. Indeed, according to the World Health Organization, about 3 million deaths are reported globally every year as a result of exposure to ambient air pollution. More alarmingly for Pakistan should be a 2014 World Bank report which states that more than 20,000 premature deaths among adults and almost 5,000,000 cases of illness among children each year are recorded in Pakistan [19] due to degraded air quality. Moreover, according to a statement available from Punjab's Environment Protection Department (EPD), the average air pollution in Pakistan's big cities is up to four times higher than the WHO limits. Steps must be taken to reduce these numbers and bring this alarming situation under control. This can only be done once the spatial distribution of aerosols is formed and regions with high concentrations of air pollution within the distribution are identified, which happens to be one of the objectives of this work. In order to build this spatial distribution, measurement of the aerosol concentration at different locations in the region of interest is required. This calls for the knowledge of air pollution measurement techniques, which will be introduced next.

2.1.5 Quantification of Pollutants

Fortunately, engineering and science has made it possible that there exist sensors and monitors that can continuously record the concentration of these pollutants and even keep a history. This monitoring is of high importance since without any

handy information extracted out of this data, no reasonable effort can be made to curb and control this crisis. Pollutants whose concentration is usually measured include Carbon Monoxide, Ozone, Sulphur Dioxide, Nitrogen Dioxide and both types of PM. The concentration of these pollutants is measured in the units of parts per million (ppm), parts per billion (ppb) and micro grams per cubic meters ($\mu g\, m^{-3}$). These concentrations are further converted into a quantity with values ranging from 0 to 500 called the Air Quality Index (AQI), which is commonly used instead of the concentrations to represent the standard of air quality. This index corresponds to five labels representing the air quality as 'good' (lower AQI) to 'hazardous' (higher AQI).

Following are the two primary ways to measure the concentration of pollutants in the atmosphere:

Ground-based Measurement Techniques:

This technique involves the installation of monitoring sensors in the desired locations to measure the concentration of pollutants in the vicinity. In Lahore, Environmental Protection Department (EPD) has installed several air quality monitors around the city, that provide daily average concentration of Sulphur Dioxide, oxides of Nitrogen, $PM_{2.5}$, PM_{10}, Carbon Monoxide and Ozone. These are located on Jail Road, Ravi Road, Band Road and Gulberg. Moreover, some privately owned sensors, measuring similar pollutants are also installed on various locations including Lahore Upper Mall, Bedian Road, Abubakr Block and NETSOL-Ghazi Road Interchange.

In addition to the above monitoring stations, AErosol RObotic NETwork (AERONET), a network of ground-based sun photometers which measures atmospheric aerosol properties is another source of ground-based measurements of

aerosol concentrations but in the form of AOT [20]. These aerosol properties are retrieved via an inversion algorithm developed by Dubovik and King in the year 2000 [21] which few years later, was further developed by Dubovik et al. to incorporate non-spherical shapes of aerosols [22]. In Pakistan, AERONET data is available at two sites, Lahore and Karachi, with the one in Lahore being operational under a collaboration of NASA and Institute of Space Technology [3].

In the context of this research, the data from ground-based sensors is of importance for the purpose of validating the satellite-based data which is discussed next.

Satellite-based Measurement Techniques:

Radiometry is a well-known technique used to measure electromagnetic radiation reflected or emitted by an object in different frequency bands. Remote sensing is based on radiometry which is of two types: active and passive sensing. Active sensing involves emission of radiation by the sensor towards the target, followed by the measurement of the reflected radiation from the target. Whereas, passive sensing, as the name suggests, involves the measurement of radiation being emitted by the target alone. Various kinds of radiometers are mounted on satellites that measure electromagnetic radiation of Earth based on these sensing principles. Literature review reveals that satellite data for conducting research on atmospheric sciences is mostly taken from MODerate resolution Imaging Spectroradiometer (MODIS) on NASA's Aqua and Terra satellites, measurements made by sensors on Landsat-8, Aura as well as European Space Agency's (ESA) Sentinel-2. In this work, data from MODIS will be used, details of which will be presented in the next section.

The major drawback of ground-based sensors is extreme sparsity of data being measured in context of spatial coverage. Due to the deployment and maintenance being expensive and cumbersome, it is impossible to install the sensor nodes almost

everywhere to get a nice uniform distribution of aerosol concentration. Fortunately, satellite-based remote sensing of aerosols provides an attractive alternative for a global coverage. It is important to note at this point that some sparsity of data with respect to spatial coverage still remains in satellite data due to some limitations (that will be revealed soon) and is indeed one of the problems that will be solved in this work via statistical modeling.

2.2 Data Description and Problem Formulation

In this section, a detailed description of dataset used will be presented first followed with the identification of potential problems.

2.2.1 MODerate resolution Imaging Spectroradiometer

MODerate resolution Imaging Spectroradiometer (MODIS) is an instrument on NASA's two satellites called Terra and Aqua. MODIS is a passive instrument which measures intensity of radiation reflected back from Earth in 36 wavelength bands (0.405 – 14.385 µm) with three spatial resolutions of 250, 500 and 1000 m. It covers the entire Earth's surface in 1-2 days, which is a reasonable temporal resolution for research under consideration. At specific bandwidth channels, different atmospheric phenomenon and particles are detected [23]. This had lead to the availability of diverse data products for land, cryosphere, ocean and atmosphere. These data products are readily available for researchers to be downloaded by NASA. Since the launch of Terra and Aqua satellites back in 1999 and 2002, respectively, MODIS data has been extensively used for atmospheric research. In this work, AOT data acquired from Aqua MODIS will be used only. Next section describes the details of how a mere radiation from Earth sensed by MODIS is converted into an entire dataset of AOT.

From Radiation to MODIS Aerosol Product:

To understand the process of how AOT dataset is developed from sensor measurements made by MODIS, it is important to look at the data processing levels of MODIS. Figure 2-2 shows a flow diagram taken from [24] that illustrates this process.

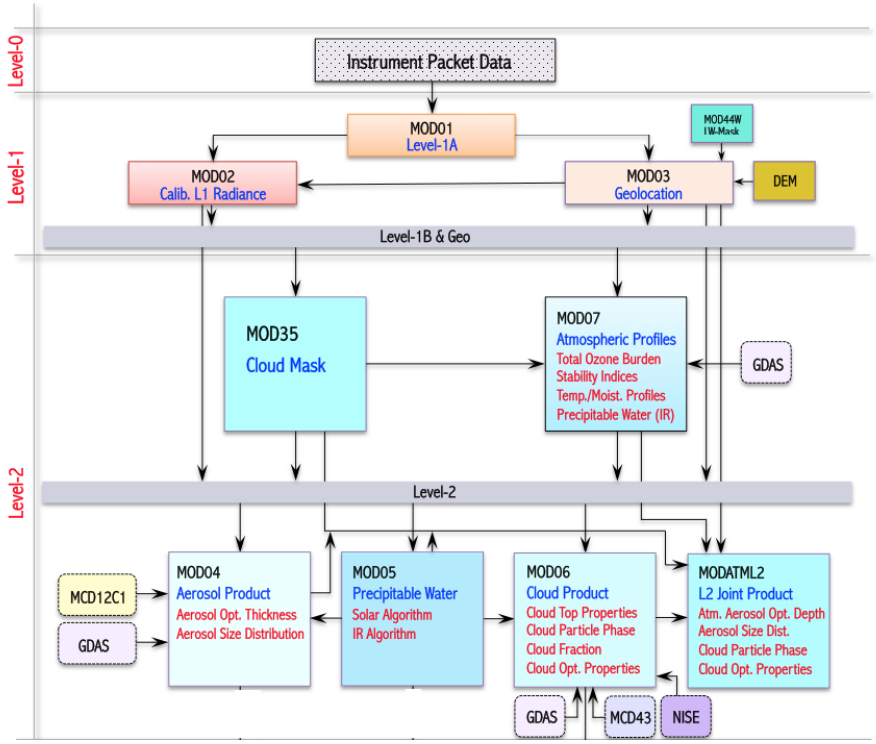

Figure 2-2: Data processing levels of MODIS.

As shown in the diagram, aerosol product is developed at the data processing level-2. Level-0, raw instrument package consisting of 5 minute swath of data, is used to produce radiance counts for each wavelength band, classified as level-1A data. These radiance counts are further converted into a more useful form, level-1B, consisting of geolocation and calibrated radiances (in SI units) for all 36 wavelength bands. Level-1B data has the spatial resolution of 250, 500 and

1000 m. However, in the context of this research, level-2 data is of much higher significance.

MODIS level-2 data is developed using information of cloud mask, atmospheric profiles along with calibrated radiances and geolocation of level-1B data. Aerosol product equipped with AOT and Aerosol Size Distribution also resides at this level with a spatial resolution of 10 km and even 3 km. AOT observed by MODIS can be thought of as the amount of light scattered or absorbed by the particles in a vertical column through the atmosphere. This, as expected, happens to be correlated with the ground-level AOT measurements made at the AERONET stations [10]. Aerosol Size Distribution is based on the Angstrom Exponent (AE) and is used to differentiate between aerosols of different dimensions. There exist various algorithms to retrieve AOT from level-1B data, most common ones of which include Dark Target, Deep Blue, Dark Dense Vegetation [25]. MODIS's 10 km AOT is based on Deep Blue and Dark Target both whereas 3 km AOT is based on Dark Target only. All the AOT retrieval algorithms work on the principle that in some particular wavelength band, surface type under observation should appear dark and aerosols should appear bright. Dark Target, for example, retrieves AOT for the locations which appear dark in visible and longer wavelengths [9]. These are usually the surfaces with vegetation and dark soil. Deep Blue considers the locations that appear dark in near ultraviolet wavelengths (hence the name Deep Blue) [8]. Deep Blue when combined with Dark Target increases the overall spatial coverage by including bright land surfaces like deserts for which Dark Target algorithm fails to retrieve AOT. Furthermore, comparison of various AOT retrieval algorithms has also been made for the benefit of researchers [26]. It is worthy to mention here that these AOT datasets inherently suffer from the problem of missing data, as described earlier. This absence of data is a consequence of MODIS not being able to observe aerosols due to, for example, cloud cover above the area

under observation.

2.2.2 Problem Formulation

There is no doubt that the first step to control the crisis of degraded air quality is to identify the locations with high concentrations of pollutants. This is exactly the problem of atmospheric pollution characterization for which, statistical and mathematical methods will serve as a backbone.

For the purpose of this characterization, it is pretty straight-forward that one would be interested to identify the locations where pollution concentration stays high for longer periods of time. These are the locations that not only suffer from high aerosol concentration but also, are responsible for dispersing these aerosols in the vicinity. Once these locations or "aerosol hot-spots" are identified, various quantification metrics can help characterize the aerosols residing in these areas. This is in fact, the first proposed idea in this work to gain more understanding of the aerosol concentration in Lahore.

However, keeping in view the limitations of MODIS dataset that is to be used i.e. it suffers from spatial coverage holes, predicting the values for these missing points using a model that governs the relationship between locations and data becomes another interesting problem to solve.

It is worthwhile to note at this point that these research dimensions have never been explored for Lahore's atmospheric data before. Although, similar other analyses and modeling has been performed based on the atmospheric data of other developing countries like China [18], [27]. Later chapters will refer to some methods used in these studies that are relevant to this work.

The next two chapters are entirely based on the theoretical details of proposed methods which were applied to achieve the objectives discerned above.

Chapter 3

Gaussian Processes for Statistical Modeling of Aerosol Optical Thickness

In past, several methods have been proposed to deal with the problem of missing AOT values in MODIS data. These methods include fusion of AOT data from sensors of multiple satellites [28], development of new AOT retrieval algorithms with better spatial coverage [29], and more recently, with the boom of machine learning and inference methods, statistical modeling techniques have also been proposed lately to solve this problem [18].

In this work, Gaussian Processes Regression (GPR), a probabilistic supervised modeling technique based on the idea of learning a model from data and using Bayesian inference to estimate the values at locations where data is missing was employed. This step of estimating missing AOT values is essential for many applications where missing data in MODIS AOT product becomes a hurdle. With this statistical model, one can readily predict the AOT value anywhere in the region under study with a certain confidence level and can infer about the relationship

of aerosol at one location with that of another location. Theoretical details of GPR will be presented in this chapter based on which, estimation and statistical modeling of AOT will be formulated in Chapter 5.

3.1 Motivation

Oftentimes, in the domain of signal processing research, there exist two or more quantities that are related to each other via a functional mapping. Some of these quantities or variables are independent which means that they do not depend on any other quantity. The others are dependent variables which take on some value according to the functional mapping involving independent quantities. In most of the cases, this mapping is unknown and could be estimated using a powerful learning technique called regression provided that the quantities involved take on continuous values. Once a suitable mapping is found that can describe this relation with a negligible error, one can readily use this mapping or model to predict values of dependent quantity at the points where they were missing. It is clearly evident that regression falls under the category of supervised learning since it requires a dataset with correspondences between the two quantities, commonly referred as the labeled data.

Regression could be of two major types - parametric and non-parametric. In parametric modeling, some assumptions on the mapping are required to be made. For instance, the assumption that a linear relation exists between the two quantities. This is a classical form of probabilistic regression known as linear regression. But this assumption of linearity will not work in every scenario. It is bound to fail if the underlying mapping is not linear (or any other degree polynomial). Moreover, in parametric techniques, one has to specify the number of parameters and exactly how they will appear in the functional mapping. This structural com-

plexity makes modeling even more specific towards one class of functions and this could lead to overfitting.

The structural specificity and enumeration of parameters can be potentially avoided by resorting to non-parametric techniques and GPR happens to be one of them. In a nutshell, other approaches of regression assume a structure of model (e.g. linear) and on the top of it, a prior distribution on the values of those parameters. GPR, in contrast, assumes a prior probability for "every possible function", with higher probabilities to the functions that are more likely to be the mapping being sought. Hence, one can use GPR to perform linear or, as a matter of fact, any kind of regression. Moreover, each prediction has an associated uncertainty or confidence level with it which gives a fair idea of how consistent the prediction is with the trained model. These are some of the characteristics that make GPR an extremely attractive and powerful regression technique.

3.2 Gaussian Processes Regression

3.2.1 Gaussian Processes

A Gaussian Process (GP) is a collection of Gaussian random variables, any finite number of which have a jointly Gaussian distribution. A GP is completely characterized by a mean and covariance function [30]. If a real process f, a function of x, is a GP with mean function $m(x)$ and covariance function $k(x_i, x_j)$, it is represented as follows:

$$f \sim \mathcal{GP}(m(x), k(x_i, x_j)), \tag{3.1}$$

where,

$$m(x) = \mathbb{E}[f(x)].$$

$$k(x_i, x_j) = \mathbb{E}[f(x_i) - m(x_i)(f(x_j) - m(x_j)].$$

For simplicity, the mean function is usually taken to be equal to zero. The covariance function $k(x_i, x_j)$, as the definition suggests, captures statistical relationship based on correlation between any two random variables $f(x_i)$ and $f(x_j)$ in the process. In GPR framework, this function is simply called the kernel, hence the notation $k(.)$. The way this function is defined, it turns out that it becomes dependent on the x values that the two random variables correspond to i.e. kernel is just a function of x_i and x_j and not of the values $f(x_i)$ and $f(x_j)$. Very often, kernels are defined as functions of the distance $\|x_i - x_j\|$ instead of other functions of x_i and x_j, making the GP a stationary process. In most of the applications, it is expected that when the distance $\|x_i - x_j\|$ is small, the two random variables $f(x_i)$ and $f(x_j)$ should be more correlated and thus, kernel should take a higher value signifying higher correlation. This is achieved by defining kernels that has a factor of $e^{-\|x_i - x_j\|}$, the squared exponential kernel is a famous example of this kind. The choice of kernel parameters further changes the way a functions looks like e.g. in squared exponential kernel, changing the length scale parameter makes the function look either wiggly or smoother [31].

The marginalization property of Gaussian distribution is the foundational backbone of GP's and its implications to carry out inference in the context of GPR are remarkably useful. This property states that if an N multi-variate Gaussian distribution is given by

$$\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}),$$

where $\mathbf{y}$ is a combination of N Gaussian random variables $\begin{bmatrix} y_1 & y_1 & ... & y_N \end{bmatrix}^T$ with $\mathbf{0}$ being an N-dimensional zero vector and $\mathbf{\Sigma}$ is an NxN covariance matrix. Then the marginalization property states that each y_i will be a univariate Gaussian random variable of the form

$$y_i \sim \mathcal{N}(0, \Sigma_{ii}),$$

with Σ_{ii} represents the variance of y_i. Similarly, any two of the random variables will form a bi-variate Gaussian distribution of the following form

$$\begin{bmatrix} y_i & y_j \end{bmatrix}^T \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma_{ij}}),$$

with $\mathbf{0}$ being a two-dimensional zero vector and

$$[\mathbf{\Sigma_{ij}}] = \begin{bmatrix} \Sigma_{ii} & \Sigma_{ij} \\ \Sigma_{ji} & \Sigma_{jj} \end{bmatrix},$$

where the diagonal entries represent the variance of each y and non-diagonal entries represent the covariance between each y_i and y_j. Here, all the Σ's will be a function of x values only as described earlier. The above idea of marginalization can be extended to more than two y's (multi-variate case) as well.

It is important to understand the role of x in the above formulation. x, in the most simplest of the applications is a 1-dimensional time-indexing variable or 2 or 3 dimensional spatial coordinates (as it will be in the case of our application). As a matter of fact, it could be any n-dimensional vector that the function under consideration (f or $\mathbf{y}$) depends on. This abstractness of the indexing variable is an attractive feature of GP framework and thus, GP assumption can be applied in many engineering problems including regression and classification.

3.2.2 Regression with Gaussian Processes

Coming back to the regression problem, assume a data $\mathcal{D}$ with N points st. $\mathcal{D} = \{\mathbf{x_i}, f(\mathbf{x_i})\}$ with $i = 1, ..., N$. Here, each data point is represented as the function value $f(\mathbf{x_i})$ corresponding to the independent quantity $\mathbf{x_i}$, which is a D-dimensional indexing vector. Then under the assumption that $f(.)$ is a GP with a certain kernel (covariance matrix) $\mathbf{K}$, the data $\mathcal{D}$ will follow the following framework:

$$\mathbf{f} \sim \ \mathcal{GP}(\mathbf{0}, \mathbf{K}),$$

with $\mathbf{f} = \begin{bmatrix} f(\mathbf{x_1}) & f(\mathbf{x_2}) & ... & f(\mathbf{x_N}) \end{bmatrix}^T$ and using the marginalization property,

$$f(\mathbf{x_i}) \sim \ \mathcal{N}(0, K_{ii}),$$

$$\begin{bmatrix} f(\mathbf{x_i}) & f(\mathbf{x_j}) \end{bmatrix}^T \sim \ \mathcal{N}(\mathbf{0}, \mathbf{K_{ij}}),$$

where

$$[\mathbf{K_{ij}}] = \begin{bmatrix} K_{ii} & K_{ij} \\ K_{ji} & K_{jj} \end{bmatrix},$$

and multiple points taken together will follow a jointly Gaussian distribution, based on the extension of the 2-dimensional case above.

Inference in Gaussian Processes Regression:

Note that in this data space of GP, there will exist points that do not belong to the data $\mathcal{D}$, corresponding to which, prediction of the function value f is required. Assume such a test point $\mathbf{x}^*$, with a corresponding f(x*) which is to be predicted. Interestingly, this $f(\mathbf{x}^*)$ together with the rest of the data points $f(\mathbf{x_i})$ for $i = 1, ...,$ N from $\mathcal{D}$ will again follow a jointly Gaussian distribution because $f(\mathbf{x}^*)$ is just another random variable in the GP. So the joint distribution will be:

$$\begin{bmatrix} f(\mathbf{x_1}) & f(\mathbf{x_2}) & ... & f(\mathbf{x_N}) & f(\mathbf{x}^*) \end{bmatrix}^T \sim \ \mathcal{N}(\mathbf{0}, \mathcal{K}),$$

with

$$[\mathcal{K}] = \begin{bmatrix} \mathbf{K} & \mathbf{K}_* \\ \mathbf{K}_*^\mathbf{T} & \mathbf{K}_{**} \end{bmatrix}.$$

Since GPR is based on Bayesian inference, values of the unknown quantities are estimated using the posterior probability i.e. $\mathbf{Pr}\left(quantity \mid observations\right)$. Therefore, the problem of predicting the value $f(\mathbf{x}^*)$ is to find the posterior probability given by $\mathbf{Pr}\left(f(\mathbf{x}^*)\,|f(\mathbf{x_1}), f(\mathbf{x_2}), ..., f(\mathbf{x_N}), \mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}\right)$. Luckily, this probability simply turns out to be the conditional probability distribution on Gaussian random variables in the GP. This can be extended from one D-dimensional test point to M D-dimensional test points. In this general case, the posterior distribution looks like $\mathbf{Pr}\left(f(\mathbf{x_1^*}), f(\mathbf{x_2^*}), ..., f(\mathbf{x_M^*})|f(\mathbf{x_1}), f(\mathbf{x_2}), ..., f(\mathbf{x_N}), \mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}, \mathbf{x_1^*}, \mathbf{x_2^*}, ..., \mathbf{x_M^*}\right)$. This conditional distribution on the Gaussian random variables follow a particular form that uses the information from the joint distribution (assuming zero mean function for simplicity) [32]:

Let's assume that the joint probability is given by:

$$\begin{bmatrix} f(\mathbf{x_1}) & f(\mathbf{x_2}) & ... & f(\mathbf{x_N}) & f(\mathbf{x_1^*}) & f(\mathbf{x_2^*}) & ... & f(\mathbf{x_M^*}) \end{bmatrix}^T \sim \mathcal{N}(\mathbf{0}, \mathcal{K}), \tag{3.2}$$

where

$$[\mathcal{K}] = \begin{bmatrix} \mathbf{K} & \mathbf{K}_* \\ \mathbf{K}_*^\mathbf{T} & \mathbf{K}_{**} \end{bmatrix},$$

with $\mathbf{K}$ is the NxN kernel matrix for the training points, $\mathbf{K}_*$ and $\mathbf{K}_*$ are NxM and MxN kernel matrices for both the training and test points, while $\mathbf{K}_{**}$ is the MxM kernel matrix for the test points only. The conditional posterior probability distribution will follow the following form:

$$f(\mathbf{x_1^*}), ..., f(\mathbf{x_M^*})|f(\mathbf{x_1}), ..., f(\mathbf{x_N}), \mathbf{x_1}, ..., \mathbf{x_N}, \mathbf{x_1^*}, ..., \mathbf{x_M^*} \sim \mathcal{N}(\boldsymbol{\mu_c}, \mathbf{K_c}), \tag{3.3}$$

where the parameters $\boldsymbol{\mu_c}$ and $\mathbf{K_c}$ are found using the kernel matrices from equation (3.2) as:

$$\boldsymbol{\mu_c} = \mathbf{K_*^T K^{-1} f(x)},$$

$$\mathbf{K_c} = \mathbf{K_{**}} - \mathbf{K_*^T K^{-1} K_*}.$$

Thus, the predictive distribution in equation (3.3) is yet another joint Gaussian distribution with mean given by the vector $\boldsymbol{\mu_c}$ and covariance matrix given by $\mathbf{K_c}$. Each test point $f(\mathbf{x_i^*})$ is a Gaussian with mean given by the i^{th} element in $\boldsymbol{\mu_c}$ and variance given by the i^{th} diagonal entry in $\mathbf{K_c}$, hence each $f(\mathbf{x_i^*})$ turned out be a Gaussian RV in the GP as was assumed earlier. In theory, the value of $f(\mathbf{x_i^*})$ that should be chosen as the prediction has to be the one that maximizes the distribution of $f(\mathbf{x_i^*})$. Since this maximizer in a Gaussian distribution is simply the mean value, the prediction for the f corresponding to each i^{th} test point will be μ_i (the i^{th} element of $\boldsymbol{\mu_c}$) with K_c^{ii} (the i^{th} diagonal entry of $\mathbf{K_c}$) uncertainty associated with the prediction. Higher the value of K_c^{ii}, more uncertain the prediction $f(\mathbf{x_i^*})$ will be and thus, less confidence will be given to the prediction.

In real-world applications, the observations in training data happen to be noisy and there is some uncertainty associated in the value that each data point takes. In sensing applications, this additive, independent and identically distributed noise is termed as the sensor noise, which, as expected, is assumed to follow a Gaussian distribution. In GPR, model for such noisy observations takes a slightly different form from what was shown above. Each data point $f(\mathbf{x_i^*})$ is assumed to be coming from the following model:

$$y_i = f(\mathbf{x_i}) + \epsilon, \tag{3.4}$$

with $i = 1, 2, ...,$ N, $\mathbf{f}$ being a GP and the sensor noise of ϵ:

$$\mathbf{f} \sim \ \mathcal{GP}(\mathbf{0}, \mathbf{K}),$$

$$\epsilon \sim \mathcal{N}(0, \sigma_n^2).$$

With this noisy and more realistic model of the observations, the joint and predictive posterior distribution derived earlier take on the following form:

$$\begin{bmatrix} y(\mathbf{x_1}) & y(\mathbf{x_2}) & ... & y(\mathbf{x_N}) & f(\mathbf{x_1^*}) & f(\mathbf{x_2^*}) & ... & f(\mathbf{x_M^*}) \end{bmatrix}^T \sim \mathcal{N}(\mathbf{0}, \mathcal{K}), \tag{3.5}$$

where

$$[\mathcal{K}] = \begin{bmatrix} \mathbf{K} + \sigma_n^2 \mathbf{I} & \mathbf{K_*} \\ \mathbf{K_*^T} & \mathbf{K_{**}} \end{bmatrix}.$$

And thus, the conditional posterior distribution will become:

$$f(\mathbf{x_1^*}), ..., f(\mathbf{x_M^*}) | y(\mathbf{x_1}), ..., y(\mathbf{x_N}), \mathbf{x_1}, ..., \mathbf{x_N}, \mathbf{x_1^*}, ..., \mathbf{x_M^*} \sim \mathcal{N}(\boldsymbol{\mu_c}, \mathbf{K_c}), \tag{3.6}$$

where

$$\boldsymbol{\mu_c} = \mathbf{K_*^T}(\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y},$$

$$\mathbf{K_c} = \mathbf{K_{**}} - \mathbf{K_*^T}(\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{K_*}.$$

Learning in Gaussian Processes Regression:

The above discussion assumes that the kernel along its parameters and the variance of sensor noise is known besides the training and test data. Essentially, it describes the method of inference in GPR after having learned the kernel and distribution of sensor noise. In reality, fitting a noisy GPR model on any data is exactly the same problem as learning the kernel and sensor noise from the training data. This is where the learning aspect of GPR comes in which addresses the questions of how to choose an appropriate kernel and variance of sensor noise. It is to be noted that the performance of prediction is highly dependent on this learning phase as

this will fix the characteristics of the model $\mathbf{y}$ given in Equation (3.4). Once the learning part is done, one can readily use the predictive distribution in Equation (3.6) for inference.

Similar to most of the other probabilistic learning methods, likelihood maximization principle is used to find the optimal parameter values in GPR. Ideally, posterior estimates for parameters should be chosen based on maximizing the posterior distribution: $\mathbf{Pr}\,(parameters \mid data)$. This is defined as a product of two distributions - likelihood function $\mathbf{Pr}\,(data \mid parameters)$ and a prior distribution $\mathbf{Pr}\,(parameters)$ according to the Bayes' Theorem. When there is little prior information known about the parameters, it is useful to maximize the likelihood function only [33]. Note that in this case, the prior distribution takes on the form of uniform probability spanning all possible values that the parameters can take. Thus, maximization of the likelihood will be equivalent to maximization of the posterior distribution. Often, the likelihood function has a pretty cumbersome form, this can be made simpler by taking its *log*. This works because *log* is a monotone which implies that maximization of the likelihood will give same results as the maximization of the log-likelihood. This log-likelihood function in GPR is defined as follows [33]:

$$\log p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = -\frac{1}{2}\mathbf{y}^T(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{y} - \frac{1}{2}\log|(\mathbf{K} + \sigma_n^2\mathbf{I})| - \frac{N}{2}\log 2\pi. \tag{3.7}$$

Note that $\boldsymbol{\theta}$ in the above equation represents all the parameters that are to be learned. The way this is employed in GPR is that a kernel is specified and then all the parameters of kernel along the variance of noise (denoted by σ_n^2) is estimated by optimizing the log-likelihood function over all the parameters. This can be achieved by applying any suitable optimization algorithm. The set of parameter values that maximize the log-likelihood function will be the optimum parameters

that should be used for inference using Equation (3.6).

Since there could be many possible kernel functions to choose from, how does one make this choice and then estimate the model parameters based on Equation (3.7)? This is where the concept of cross-validation comes into play. In this process, several kernels that could most likely represent the data are chosen and optimal parameters based on Equation (3.7) are found for each of them. At this point, there will be several models similar to the one defined in Equation (3.4) at hand, each based on one of the chosen kernels and the corresponding optimum parameters. Now to choose a model among these, k-folds cross validation technique can be employed on each of these and the model that gives the best performance (based on e.g., the mean squared error between actual and predicted values) could be a potential model for the data under consideration.

This concludes the discussion of theoretical details of GPR. These concepts from GPR theory will be used in Chapter 5 to formulate and resolve the problem of AOT modeling. This proposed idea will not only solve the problem of missing data in AOT product of MODIS but will also provide a statistical model that will be able to predict AOT value with a certain confidence level at any location in Lahore, making it a much more significant result of this work.

Next chapter will discuss the theoretical concepts needed to undertake the second potential research dimension of aerosol characterization in Lahore.

Chapter 4

Characterization of Aerosol Hot-spots using Glowworm Swarm Optimization Algorithm

The objective of this chapter is to introduce the theory of proposed mathematical method which will be employed for the spatio-temporal characterization of pollution hot-spots based on AOT data. Hot-spots in this text are considered to be those locations where the AOT values persistently remain high relative to their surroundings.

For the purpose of characterization, hot-spot locations are needed to be identified first. To identify these locations, a suitable optimization method called Glowworm Swarm Optimization (GSO) algorithm is proposed which was developed in 2005 by Krishnanand and Ghose [34]. It is one of the most relevant optimization algorithms for the application under consideration. The reasons for choosing GSO over other algorithms will shortly become clear in the next section where the distinctive features of algorithm are described followed by its working principle.

4.1 Motivation

As mentioned earlier, the most significant goal in the spatio-temporal characterization is to first identify the hot-spot locations in the region based on MODIS AOT data. Interestingly, this problem is exactly equivalent to the problem of locating all the maxima of the daily MODIS data grid. In more technical terms, the problem is to locate all the local maxima of a two-dimensional AOT discrete multi-modal signal.

These kind of problems where the locations corresponding to the maxima of a function are required to be found fall under the category of optimization problems. Since the objective function in this study is a multi-modal function, an optimization algorithm that can locate all local maxima of the function should be considered.

Multi-modal functions usually model signals with multiple signal sources like sound, heat, light and leaks in pressurized systems [35]. They model well the source profiles of such signals that originate from a point and spread in the surroundings. Intuitively, when these signals are optimized, the locations from where the signal originated are found. Population-based approaches are well-suited for the multi-modal optimization and are divided into two main categories - Evolution Computation (EC) and Swarm Intelligence (SI) techniques [35]. EC techniques are based on evolutionary mechanism encountered in natural selection are not very relevant in the context of aerosol hot-spot identification, leaving behind the SI algorithms.

SI algorithms are meta-heuristic in nature which means that they are based on high level methods that perform sufficiently good for optimization problems. The word swarm is used for a dense group, usually that of small biological creatures. Hence, SI algorithms consist of agents that locally interact with one another and their environment, mimicking the behavior of biological systems like

ant colonies and bacterial growth. Some of the well-known SI algorithms include Particle Swarm Optimization (PSO), Ant Colony System (ACS) and Artificial Bee Colony (ABC) [36]. Similar to these algorithms is Glowworm Swarm Optimization (GSO) algorithm, which, as the name suggests, mimics the behavior of glowworms.

GSO is different from the earlier approaches to multi-modal optimization in the dynamic decision domain that the agents in the swarm use to locate multiple peaks. Moreover, it is memoryless, gradient free and doesn't even require knowledge of global information of the signal to be optimized [35]. These features of GSO makes it a reasonable choice to work with. As a matter of fact, there exists a recent experimental study in which, after slight modifications in the GSO algorithm, back-scattering of the aerosols in Chengdu city of China is modeled to identify the air pollution sources [27]. However, the use of GSO in this study is to only locate the hot-spots which will be the first step for the aerosol hot-spot characterization of Lahore. Next section is devoted to elaborate the working principle of GSO algorithm.

4.2 Glowworm Swarm Optimization (GSO) Algorithm

Being an SI algorithm, GSO algorithm is based on a population of agents that interact with one another and their environment to optimize the objective function. Each agent in GSO algorithm is called a glowworm. These artificial glowworms possess characteristics similar to their natural counterparts. This includes a Luciferin-level associated with each glowworm in GSO which is analogous to the glow in natural glowworms. This glow helps the glowworms to interact better with each other and their environment. Similar benefits can be achieved from the artificial glowworms in GSO algorithm by exploiting this Luciferin-level. Like many

other optimization methods, GSO algorithm can be employed for seeking either the local maxima or the local minima by making slight changes in the algorithm. Since in this thesis, positions of local maxima are of interest, the version of GSO used to locate local maxima will be discussed. Next section introduces three basic mechanisms that govern the whole working principle of GSO algorithm.

4.2.1 Three Mechanisms of GSO Algorithm

In GSO, all the glowworms are given a default Luciferin-level and are randomly spread out in the search space of the objective function. How the glowworms will move around and interact with one another to locate the local maxima of the objective function rely on the following three mechanisms [34]:

Fitness Broadcast:

The Luciferin-level associated with each glowworm takes its value based on the fitness of its location in the search space. It is denoted by l_i for every i^{th}-glowworm. Intuitively, the closer a glowworm is to a local maxima, the higher the value of its Luciferin-level. This value is broadcasted to all glowworms in the search space making each glowworm know the Lucifern-level of every other glowworm despite large distances among them. Note that this capability of unconstrained perception is not possessed by the natural glowworms and it tends to reduce as the distance among them increases.

Adaptive Neighborhood:

Each glowworm involved in the algorithm has a local decision range. This, as mentioned earlier, is a dynamic or adaptive range which is used to form a neighborhood for every glowworm. It is based on a variable range denoted by r_d^i which is bounded by a hard-limited sensor range r_s, such that $0 < r_d^i < r_s$. r_d^i is modulated

based at every iteration to ensure the adaptiveness of decision range. Based on this r_d^i, a glowworm forms its neighborhood range and chooses those glowworms as its neighbors that have higher Luciferin-level than itself.

It is important to understand the significance of this feature. In the multimodal functions where the local maxima are being sought, agents use their vision (decision range) to decide where the peak is and move towards that direction. If this range is fixed, it has to be smaller than the minimum distance between the peaks in the function, otherwise, some peaks will not get located. Since little information is known about the function to be optimized, it is better to have a dynamic decision range which will update itself based on the information of local neighborhood. This is exactly what GSO uses to locate all the peaks.

Positive Taxis:

Among all the neighbors of a glowworm, there will be one neighbor that the glowworm will choose to move towards. Since the glowworms move towards the stimulus (higher Luciferin-level), this movement is termed as the positive taxis. The selection of this neighbor among all the others is based on some probabilistic heuristic.

The above three mechanisms interplay with one another to form the entire algorithm of GSO, which will be revealed next.

4.2.2 Working Principle of GSO

The GSO algorithm consists of four main steps, three of which are repeated until the local maxima in the objective function are found. The description of these steps follow next.

Step 1: Initialization

In this step, N glowworms are randomly spread out in the search space, each initialized with a default Luciferin-level l_0, regardless of their position. Other parameters that govern the GSO algorithm are also initialized in this step.

Step 2: Luciferin Update

After initialization, each i^{th}-glowworm's Luciferin-level will be updated based on its position in the search space. This value, as described earlier, is proportional to the value of the objective function at the position of the glowworm. The Luciferin-level of an i^{th}-glowworm will be updated based on the following equation:

$$l_i(t+1) = (1-\rho)l_i(t) + \gamma J(x_i(t+1)). \tag{4.1}$$

Here, t represents the iteration number, $J(.)$ is the objective function that is to be optimized possessing multiple maxima, ρ is the Luciferin decay constant $0 < \rho < 1$ which controls how much of the previous Luciferin-value has to be retained and γ is Luciferin enhancement constant which scales the Luciferin-level with the function value.

Step 3: Movement Phase

After all the glowworms attain an appropriate Luciferin-level based on the fitness of their position, they are required to move towards a local maxima. This is achieved by the glowworms by moving towards a neighbor that has a higher Luciferin-level based on a probabilistic heuristic. The heuristic represents the probability with which an i^{th}-glowworm will move towards a j^{th}-glowworm and is defined as:

$$p_{ij}(t) = \frac{l_j(t) - l_i(t)}{\sum_{k \in N_i(t)} l_k(t) - l_i(t)}, \tag{4.2}$$

where j belongs to $N_i(t)$ with $N_i(t) = \{j : d_{ij}(t) < r_d^i; l_i(t) < lj(t)\}$ being a set of neighbors of a glowworm i that glow brighter than the itself and are located at a Euclidean distance of less than r_d^i, the adaptive decision range.

This is how each i^{th}-glowworm in the search space forms a neighborhood $N_i(t)$ and moves towards a j^{th}-glowworm with a probability of $p_{ij}(t)$. At iteration number t, the update equation for the position of each glowworm is given by:

$$x_i(t+1) = x_i(t) + s\frac{x_j(t) - x_i(t)}{\|x_j(t) - x_i(t)\|}. \tag{4.3}$$

In the above equation, $x_i(t) \in \mathbb{R}^m$, represents the position of the i^{th}-glowworm at iteration t in an m-dimensional space, $\|.\|$ is the Euclidean norm operator and s is a positive quantity that denotes the step size.

Step 4: Neighborhood Range Update

This is the last step in each iteration of the algorithm. After all the glowworms have moved towards their chosen neighbors according to the above two equations, they are all at a new position now (which is more closer to a local maxima). At this point, due to the adaptive decision range property of GSO, each glowworm needs to update the value of r_d^i for forming a neighborhood in the next iteration. The decision range of an i^{th}-glowworm is updated based on the following rule:

$$r_d^i(t+1) = \min_{r_s}\{r_s, \max\{0, r_d^i(t) + \beta(n_t - N_i(t))\}\}. \tag{4.4}$$

Here, $r_d^i(0)$ will be initialized at the first iteration as r_0 for all i. r_s is a fixed sensor range which bounds the $r_d^i(t)$, nt controls the number of neighbors, $N_i(t)$ measures total number of neighbors of the glowworm i in the current iteration.

After conducting extensive numerical experiments, appropriate values of most of the parameters in GSO algorithm have been determined [33]. These include n_t, r_s, l_0, β, ρ, r_0 and γ and their values are tabulated below. While the remaining two parameters N and r_s control the number of peaks that are captured by the algorithm. These two parameters can be selected according to the requirement of the application.

ρ	γ	β	n_t	s	l_0
0.4	0.6	0.08	5	0.03	5

This concludes the discussion on the theory of GSO algorithm which will be employed to identify aerosol hot-spots based on MODIS AOT data, the details of which will be presented in the next chapter.

Chapter 5

Experiments and Results

In this chapter, the identified problems will be formulated formally based on the methods described in the Chapters 3 and 4. Study site description, useful notation and some preliminary analyses carried out on the data will be presented first followed by all the experiments that were conducted to achieve the objectives of this thesis with their associated findings and conclusions.

5.1 Study Site Description

Lahore is one of the biggest cities of Pakistan, situated in the province of Punjab. Increased aerosol concentration in Lahore is a consequence of city's vehicular and industrial emissions, biomass burning activities and dust aerosols [3]. Figure 5-1 shows the area around Lahore for which all the analyses were carried out (highlighted with red rectangular boundary). It consists of the city of Lahore, small towns in the outskirts, some field areas and several highways that connect Lahore to other cities.

Figure 5-1: Region under study (Image courtesy: Google Maps).

5.2 Data Exploration and Notation

In this study, the aerosol product of Aqua MODIS available in the spatial resolution of 3 km, the details of which were given in Section 2.2.1, was used. More specifically, this comprises of the data product named as MYD04_3K, associated with the years 2017 and 2018. This data was downloaded from the online LAADS DAAC data archive [37]. Although, the 3 km product generally happens to have less spatial coverage due to the fact that it uses Dark Target algorithm instead of Deep Blue, it works good enough for the dark surface of Lahore. Moreover, the choice of 3 km product was made because only the region in and around Lahore is

of interest, so it is better to use a finer resolution to ensure that more data points can cover the small region of Lahore. Next, the notation for this AOT data will be developed to facilitate the understanding of all the components that are involved in the dataset. This notation will be extensively used to describe the problem formulation, methodology and relevant results in the later sections.

5.2.1 Notation

Let the AOT data points for a d^{th} day of a particular year be denoted as $\mathcal{D}^d$. st. $\mathcal{D}^d = \{\mathbf{x_i}, y_i\}$. Here, $d = 1, ..., 365$ represents the day of year and $i = 1, ..., N$, the data point number, with N being the total number of data points available in $\mathcal{D}^d$. $\mathbf{x_i}$ is a two dimensional vector representing the geographical coordinates with x_1 as the latitude and x_2, the longitude. In this study, $\mathbf{x_i}$, the geographical coordinates will always belong to the region of Lahore and its outskirts, denoted as $\mathcal{R}_\mathcal{L}$, where $\mathcal{R}_\mathcal{L} \triangleq \{\mathbf{x} \mid 31.2 \leq x_1 \leq 31.7, 74 \leq x_2 \leq 74.5\}$. y_i denotes the value of AOT observed by MODIS at the coordinates defined by $\mathbf{x_i}$.

For the data points in the daily AOT data for which the observation is missing corresponding to some $\mathbf{x_i}$, a new notation $\mathbf{x_i^*}$ with $i = 1, 2, ..., M$ will be used.

5.3 Preliminary Data Analyses

In this section, some elementary analyses that were carried out on the data before conducting the actual research are presented. These experiments were performed to ensure that the datasets are valid and follow similar trends that were observed in the studies before. Two major types of analyses were conducted for this purpose: correlation studies between satellite and ground-based data, and seasonal trends of AOT. The details of these analyses are presented next.

5.3.1 Analysis I: Correlation between MODIS AOT and AERONET AOT

For the purpose of validating the accuracy of satellite data, it is important to establish a positive correlation between the satellite and ground-based measurements before conducting any research using this data. This will ensure that the satellite data is consistent to what is observed by the sensors on ground and thus, is suitable to use for conducting research. Although, it has already been verified in previous studies that the MODIS AOT data is positively correlated with the AOT measurements of AERONET ground stations, a similar study for determining the correlation of satellite AOT with AERONET AOT was conducted in this work for the purpose of validating MODIS data.

Correlation between two quantities is usually determined using a scatter plot. In a scatter plot, each of the two quantities is plotted on one of the two axes. Each point on the plot corresponds to a measurement taken at the same time-stamp (day of year) and at the same location for both quantities.

To find correlation between MODIS AOT and AERONET AOT, MODIS AOT value is required corresponding to every ground-based measurement by AERONET. As discussed in the earlier chapters, MODIS dataset suffers from the problem of data loss, due to which, ground-based measurements will not have their corresponding satellite measurements on all the available time-stamps. On the days when MODIS AOT is available on the location of the corresponding ground-based measurement, the value of the MODIS AOT around the coordinates of AERONET location is averaged to find the corresponding MODIS AOT value for correlation. Using this method, correlation analysis was carried out, the relevant discussion and results follow next.

Results and Discussion:

In this experiment, AOT from AERONET's daily averaged cloud screened (level 1.5) data of two years (2017-2018) was used to validate the MODIS AOT data. This data was downloaded using the AERONET Data Download Tool [38]. Since AERONET doesn't measure AOT in 550 nm wavelength band, it was required to interpolate the value of $\text{AOT}_{550\text{nm}}$ from $\text{AOT}_{500\text{nm}}$, which is provided in the data. This was calculated using a relationship based on power rule which is as follows [4]:

$$AOD_{550\text{nm}} = AOD_{500\text{nm}}(\frac{550}{500})^{-\alpha},$$

where α represents the Angstrom Exponent at the wavelength of 440-870 nm measured at the AERONET station. Figure 5-2 shows scatter plots between the Aqua MODIS AOT and the AERONET AOT for the two years. N represents the total number of match-ups that were found, R-square is the coefficient of determination (R^2) that describes how well a linear model fits this data. From the figures, R^2 for the year 2017 doesn't suggest a very good fit as compared to the one for 2018. The reason for this was not explored in this work. However, the plot for year 2018 suggests reasonable positive correlation between the two quantities, making the result of this analysis consistent with the previous studies.

5.3.2 Analysis II: Seasonal Variations in Aerosol Optical Thickness

This is the second analysis that was done on the AOT data to understand how the concentration of aerosols in a region varies with time. This analysis has been often conducted in the previous studies, as described in the introductory chapter.

In this analysis, the average value of AOT was calculated for each day using the available data points in the region of Lahore, $\mathcal{R}_L$. To find the average AOT

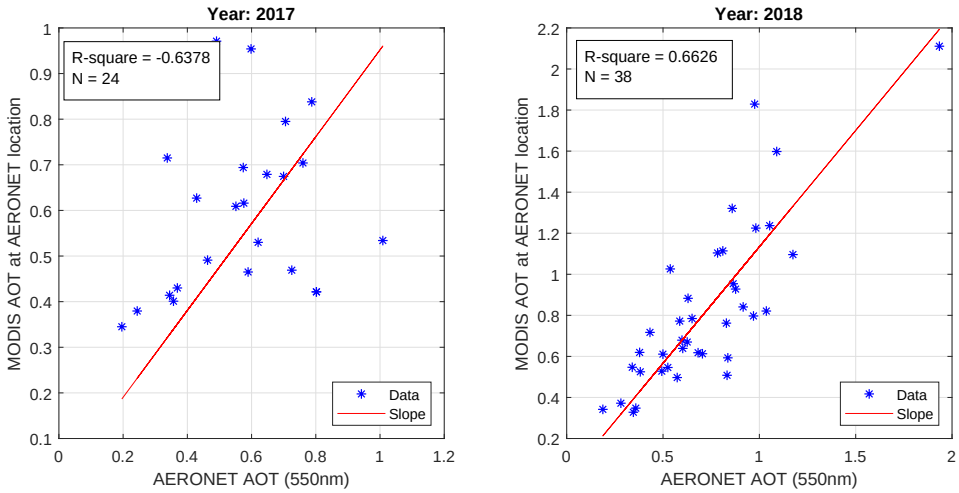

Figure 5-2: Correlation between MODIS AOT and AERONET AOT.

value for each season - Winter, Pre-Monsoon, Monsoon and Post-Monsoon, AOT values were further averaged according to the months that fall in each season. The seasons were defined in terms of months as follows:

- Winter: December, January, February
- Pre-Monsoon: April, May, June
- Monsoon: July, August, September
- Post-Monsoon: October, November

Results and Discussion:

Figure 5-3 shows the variation of AOT values found for the two years, 2017 and 2018. Clearly, Monsoon season has the highest concentration of aerosols, followed by the two seasons of Pre and Post Monsoon, with the lowest value observed for Winter. The reason for the highest AOT value for the Monsoon season is usually associated with the frequent dust storms that produce dust particles in the atmosphere, increasing the concentration of dust PM in the air. This result also happens to be in consistency with the previous studies.

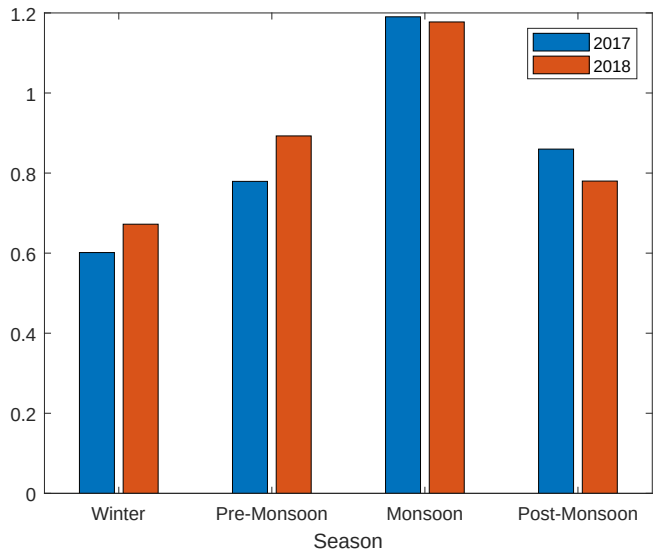

Figure 5-3: Seasonal variation in Aerosol Optical Thickness.

This concludes the discussion on the two analyses that were conducted on AOT data of Aqua MODIS. The analyses discussed above suggest the validity of MODIS AOT data, and the results, being consistent with the previous studies, propose that MODIS AOT data can be used for the purpose of research. Using MODIS data, the experiments involving statistical modeling of AOT over Lahore will be described next.

5.4 Statistical Modeling of AOT

In this section, experimental details of learning a statistical model of AOT over Lahore are presented. Recall that this was the proposed solution for dealing with the problem of missing AOT data observed by MODIS. First off, all the theoretical assumptions that were made on the data will be described followed by the methodology of learning the statistical model based on those assumptions. These assumptions, based on Gaussian Processes Regression framework (introduced in

Chapter 3), will be used to formulate and carry out the learning of model. All the results and findings related to the learned model will be presented. Since, using the learned model, inference was also carried out for the coordinates where MODIS AOT data was missing, the inference method along with its associated results will be shown as well.

5.4.1 Model Assumptions

To determine the statistical relationship between the coordinates of Lahore and AOT values for a given dataset $\mathcal{D}^d$, it was assumed that AOT at each location in the region of interest $\mathcal{R}_L$, follows a univariate Gaussian distribution. This, in turn means that any multiple AOT points taken together will follow a multivariate Gaussian distribution with a certain mean vector and a covariance matrix (kernel) defined by the coordinates $\mathbf{x}_i$. Moreover, if it is assumed that the observations of AOT from MODIS are noisy, i.e., each AOT value has an identical and independent additive Gaussian noise term ϵ, then the model for observations in a certain $\mathcal{D}^d$ becomes exactly the one that was defined in Equation (3.4). Formally, the model for each AOT observation can be defined as follows:

$$y_i = AOT(\mathbf{x_i}) + \epsilon, \tag{5.1}$$

with $i = 1, 2, ..., N$, with all of the N data points in **AOT** collectively forming a GP based on the univariate Gaussian assumption. Hence, **AOT** and the additive sensor noise of ϵ can be represented as:

$$\mathbf{AOT} \sim \ \mathcal{GP}(\mathbf{0}, \mathbf{K}).$$

$$\epsilon \sim \ \mathcal{N}(0, \sigma_n^2).$$

In the learning of statistical modeling of AOT, it is this kernel $\mathbf{K}$ along with its hyperparameters (depending on the choice of kernel) and the sensor noise variance σ_n^2, which are required to be found. Next section is devoted to describe the methodology of learning these components that completely define the statistical model of AOT.

5.4.2 Learning of Statistical Model:

As elaborated in the Section 3.2.2, learning a statistical model based on the Gaussian Process (GP) assumption requires the maximization of the likelihood function, or the log-likelihood function given by the following equation:

$$\log p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = -\frac{1}{2}\mathbf{y}^T(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{y} - \frac{1}{2}\log|(\mathbf{K} + \sigma_n^2\mathbf{I})| - \frac{N}{2}\log 2\pi. \tag{5.2}$$

Similarly, for learning the statistical model of AOT, based on the GP assumption and defined in Equation (5.1), the above equation has to be maximized with respect to all parameters $\boldsymbol{\theta}$ (consisting of the kernel parameters and sensor noise variance). This log-likelihood maximization requires the available data points $\mathbf{y}$, which were made equal to all available AOT data points from $\mathcal{D}^d$ for each d separately, where as N represents the total number of available data points in $\mathcal{D}^d$. It should be noted here that some days do not contain a lot of data points making it difficult to learn a model from it and therefore, no model was learned for those days. As for the kernel $\mathbf{K}$, all possible kernels can be searched over to see which one of those maximize the log-likelihood function the best. This method was carried out in MATLAB with $\mathbf{K}$'s as all the available built-in kernels in MATLAB. As a result of this experiment, the ARD Exponential kernel turned out to be the one that optimized the log-likelihood function for the highest number of days.

This experiment suggests that ARD Exponential kernel could potentially ex-

plain the relationship between coordinates of Lahore and MODIS AOT. Backed by this result, $\mathbf{K}$ in the log-likelihood function was fixed as ARD Exponential and maximization was carried out again with respect to the parameters $\boldsymbol{\theta}$, for each dataset $\mathcal{D}^d$, giving d set of parameters: $\boldsymbol{\theta_d}$ for each day. This experiment revealed the most potential values of parameters that can completely characterize the kernel ARD Exponential and variance of sensor noise, leading to the formulation of statistical model for AOT.

ARD Exponential Kernel:

The statistical model of AOT found in this experiment heavily relies on the ARD Exponential (ARDE) kernel that governs the relationship of one AOT value with another in the region $\mathcal{R}_{\mathcal{L}}$. ARD (Automatic Relevance Determination) Exponential kernel is a stationary kernel, a function of $(\mathbf{x} - \mathbf{x}')$ and thus, invariant to translation in the input space [30]. Interestingly, it is closely related to the well-known Squared Exponential (SE) kernel. To gain a better understanding of the characteristics of the ARDE kernel, comparison of the definition of SE and ARDE kernel can be made since the structure of SE kernel is easy to interpret. The SE kernel is defined as:

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\bigg(-\frac{1}{2}(\mathbf{x} - \mathbf{x}')^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \mathbf{x}')\bigg),$$

with $\boldsymbol{\Sigma}$ as the DxD covariance matrix and σ_f^2 is a scaling factor, present in every kernel that determines the average distance of the function away from its mean [39]. If $\boldsymbol{\Sigma}$ is made a diagonal matrix of the form $\text{diag}(\sigma_1^{-2}\sigma_2^{-2}..., \sigma_D^{-2})$, the SE kernel becomes an ARDE kernel, defined as follows:

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\bigg(-\frac{1}{2}\sum_{j=1}^{D} \frac{1}{\sigma_j^2}(x_j - x_j')^2\bigg),$$

In the context of statistical model of AOT, structure of ARDE kernel suggests that the correlation between any two AOT points in data will depend on the weighted Euclidean distance, weighted by the length scale parameters given by σ_j^2, with j representing the dimension; as opposed to the more general Mahalanobis distance in the SE kernel.

Based on the entire discussion above, it can be concluded that for the AOT model in Equation (5.1), with $\mathbf{K}$ as the ARDE kernel, there will be a total of four learned parameters: three hyperparameters of ARDE kernel i.e., two length scales σ_1^2 and σ_2^2 corresponding to the latitude and longitude and a signal standard deviation σ_f^2, and one variance parameter σ_n^2 from the model equation. Recall that the statistical model was fit for every day of the year, this will provide us with a set of these four parameters (e.g. 365 values per each parameter for one year). The variations seen in the values of these parameters for the year 2017 across days is shown in the Figure 5-4. Moreover, the associated histograms for these time-series are presented in Figure 5-5.

These plots suggest that the model is not a stationary model across time since there is a reasonable amount of variation in the values. Similarly, parameters for year 2018 were also found to be non-stationary with time. Interestingly, it was observed that the lengthscale parameter was taking unreasonably larger values for certain days. To learn more about what was causing this behavior, data belonging to these days was explored further. As expected, this anomaly was found to correspond to the days when small number of points were available (refer to Figure 5-6) or when data is very densely present at a particular location and missing on all the other locations as shown in Figure 5-7. This kind of data is not suitable to build a statistical model. Note that the figures illustrating the parameter values were plotted after removing these outliers for the ease of data interpretation. One possible solution to deal with this problem could be to employ a temporal model

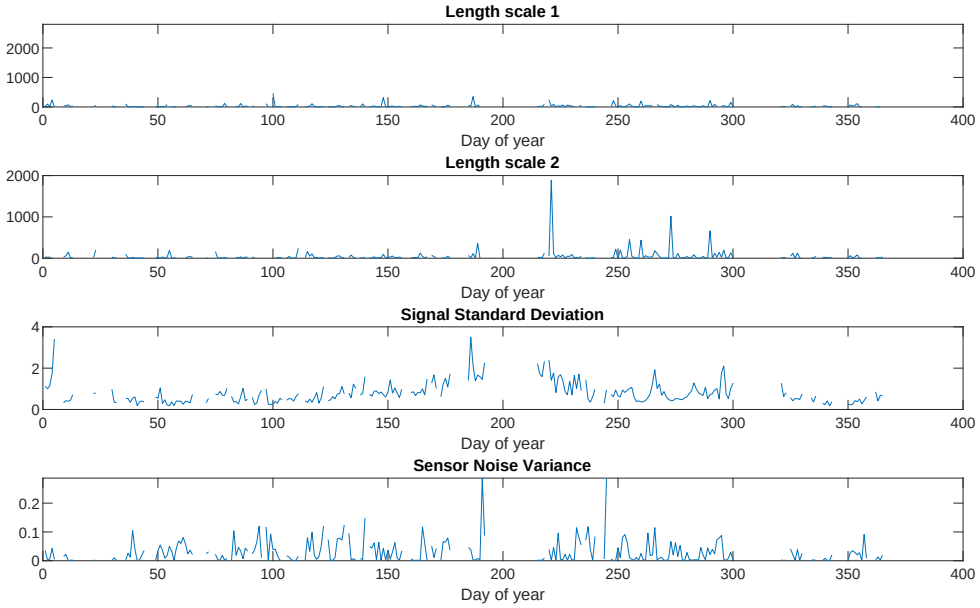

Figure 5-4: Time-series of parameter values (Year 2017).

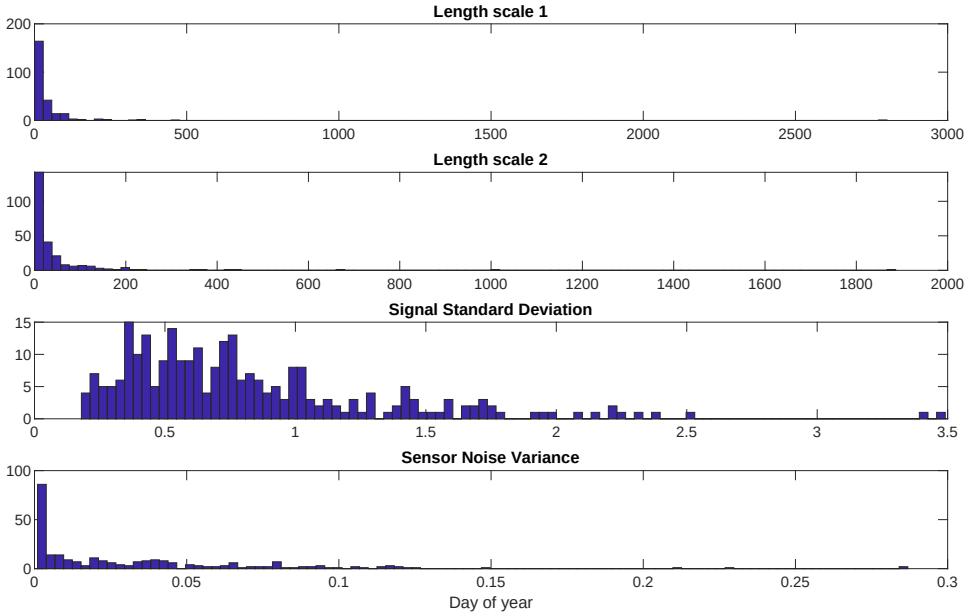

Figure 5-5: Histogram of parameter values (Year 2017).

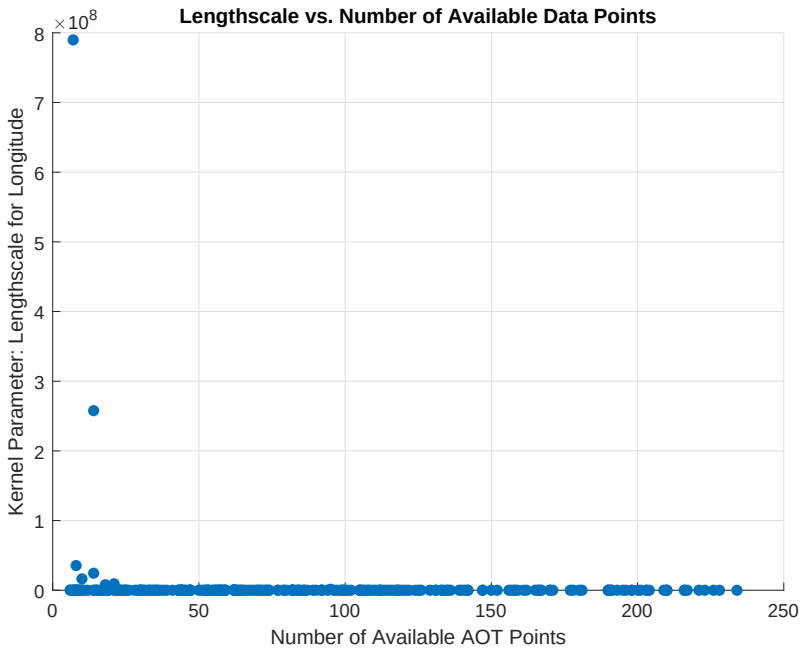

Figure 5-6: Scatter plot between lengthscale of longitude and number of available AOT points.

that can use data from the previous and the next day to interpolate these missing points. However, in this work, spatial model is proposed only and thus, this is one limitation of the proposed model. Next section addresses about the evaluation of this learned model.

Evaluation of the Learned Statistical Model:

To evaluate how well the learned model represents MODIS AOT data, 10-folds cross-validation technique was used. This method splits the data into 10 folds, considering the 9 of those parts as the training data and the last one as the test data. This division is done 10 times, each time, generating the parts using random data points. Each time, an ARDE kenrel-based GP model was fit on the 9 folds of training data and the AOT values were predicted for the points in the tenth fold.

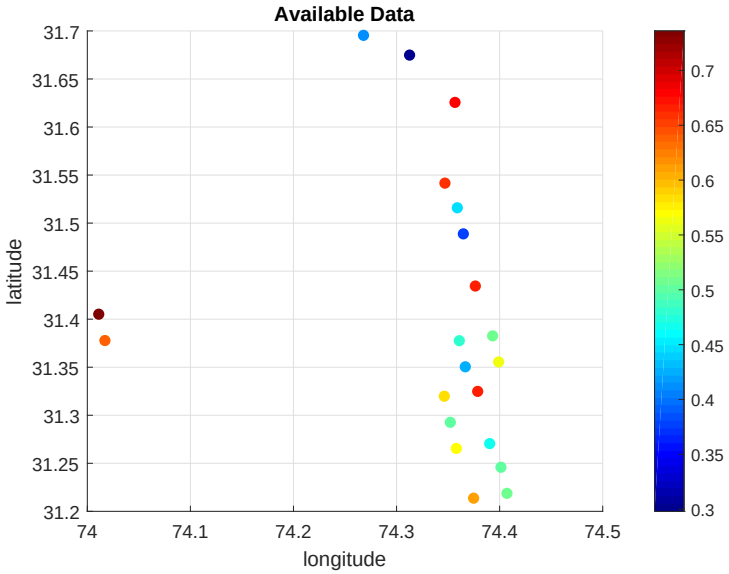

Figure 5-7: An instance of data corresponding to large value of lengthscale parameters.

Since the true AOT values form the data is already known, error can be computed by comparing the difference between the actual and predicted AOT value. An averaged error corresponding to all the 10 divisions of data can be computed as well which is termed as the cross validation or CV error.

Using 10-folds cross validation for the model of each day, CV error based on two error metrics was computed which reflected the fitness of the learned model for the data. The two metrics that were used in this work are Mean Squared Error (MSE) and normalized MSE (nMSE). However, the second metric, nMSE, is a better metric since it determines the error by incorporating the energy of actual values of AOT as a normalization factor. The definition of the two evaluation metrics follow next.

- Mean Squared Error:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2.$$

- Normalized Mean Squared Error:

$$nMSE = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i)^2},$$

where $\hat{y}_i$ represents the predicted AOT value corresponding to the i^{th} test data point and y_i is the actual AOT value of the i^{th} test data point.

The cross-validation MSE and nMSE for the learned model was determined using the above methodology. The CV errors were further averaged over the four seasons to interpret the error values easily.

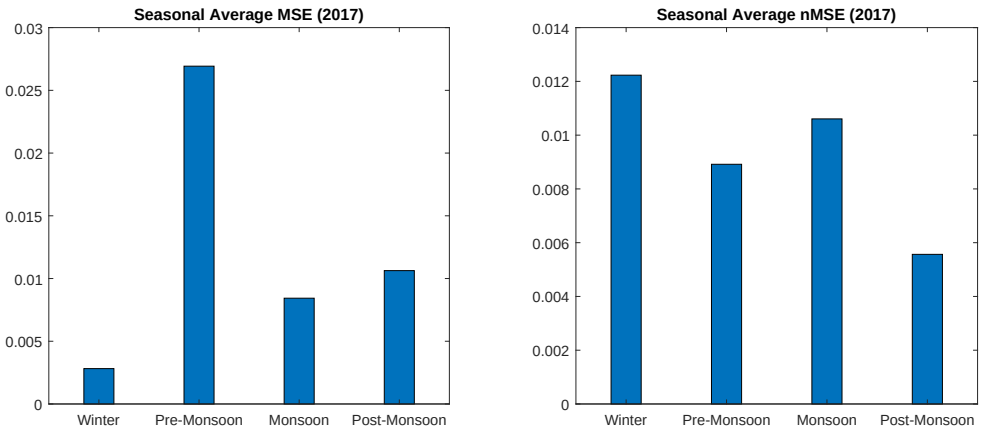

Figure 5-8: Seasonal averaged 10-folds cross validation errors for the year 2017.

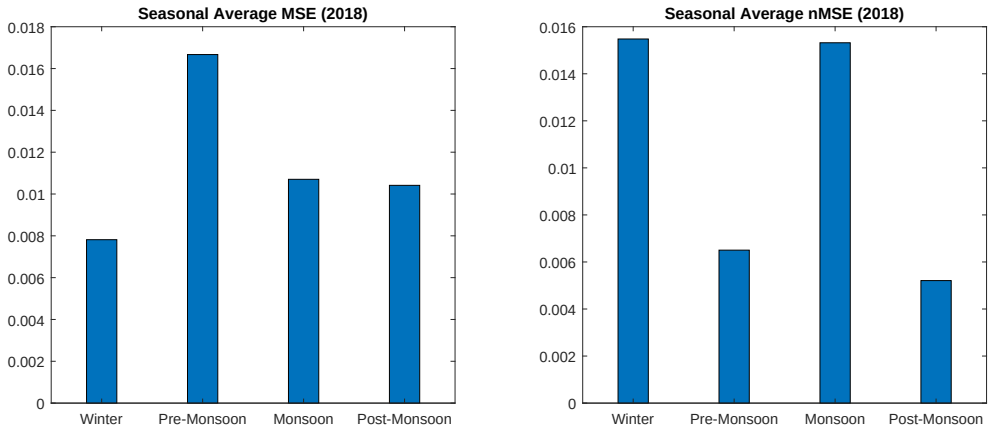

Figure 5-9: Seasonal averaged 10-folds cross validation errors for the year 2018.

Figures 5-8 and 5-9 illustrate these errors on bar plots for each season. These further averaged to yearly CV nMSE becomes 0.0093 and 0.0106 for 2017 and 2018 respectively. It can be easily observed that the error values are reasonably good. Particularly the values of nMSE strongly suggest that the learned statistical model is a potential model for AOT for the region $\mathcal{R}_{\mathcal{L}}$, representing Lahore. As mentioned earlier, a utility of this model is that one can infer about value of AOT anywhere in the region $\mathcal{R}_{\mathcal{L}}$ i.e. prediction of the missing values can be done. The next section represents the method to carry out inference using this model along with some relevant results.

5.4.3 Inference: Predicting the Missing Values with the Learned Statistical Model

Based on the model assumption of AOT, it is not hard to realize that the problem of predicting the missing values in MODIS AOT data can be solved using the Equation (3.6). The formulation of equations for inference of AOT will take the

following form:

$$\begin{bmatrix} y(\mathbf{x_1}) & y(\mathbf{x_2}) & ... & y(\mathbf{x_N}) & AOT(\mathbf{x_1^*}) & AOT(\mathbf{x_2^*}) & ... & AOT(\mathbf{x_M^*}) \end{bmatrix}^T \sim \mathcal{N}(\mathbf{0}, \mathcal{K}), \tag{5.3}$$

where

$$[\mathcal{K}] = \begin{bmatrix} \mathbf{K} + \sigma_n^2 \mathbf{I} & \mathbf{K_*} \\ \mathbf{K_*^T} & \mathbf{K_{**}} \end{bmatrix}.$$

And thus, the conditional posterior distribution for making the predictions is given as:

$$AOT(\mathbf{x_1^*}), ..., AOT(\mathbf{x_M^*}) | y(\mathbf{x_1}), ..., y(\mathbf{x_N}), \mathbf{x_1}, ..., \mathbf{x_N}, \mathbf{x_1^*}, ..., \mathbf{x_M^*} \sim \mathcal{N}(\boldsymbol{\mu_c}, \mathbf{K_c}), \tag{5.4}$$

where

$$\boldsymbol{\mu_c} = \mathbf{K_*^T}(\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y}.$$

$$\mathbf{K_c} = \mathbf{K_{**}} - \mathbf{K_*^T}(\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{K_*}.$$

Here each i^{th} missing AOT data point for the coordinates $\mathbf{x_i^*}$ takes the form of a Gaussian distribution, with the predicted value given by the i^{th} element of $\mu_\mathbf{c}$ and the uncertainty in prediction given by the i^{th} diagonal entry of $\mathbf{K_c}$, as elaborated in the Section 3.2.2. Moreover, inference based on the Equation (5.4) was made using the exact method, as opposed to approximation methods which are used if the number of data points in a dataset are very large. [40].

Note that in the above formulation, $\mathbf{K}$ is nothing but the ARDE kernel with the learned parameters and σ_n^2 is the learned variance of the additive sensor noise, shown in Figure 5-4. One value out of the time-series will be chosen depending on the day for which the missing value is being predicted. An important point to note here is that inference can only be made for the data points that belong

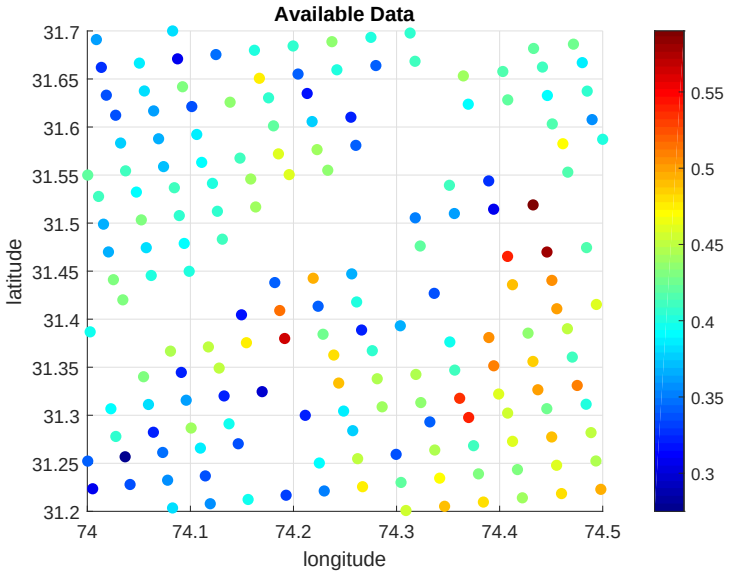

Figure 5-10: Missing AOT values in $\mathcal{D}^{19}$, year 2017.

to some $\mathcal{D}^d$ with reasonable number of points available to be used as the training data, because models were learned only for those days.

In Figure 5-10, AOT data for the year 2017 and $\mathcal{D}^{19}$ is shown. The figure clearly shows that there is a significant number of missing points in this data. Figure 5-11 shows the result of using the above method of inference to predict the AOT values for the same day. How reliable are these predictions? Figure 5-12 answers this question as it depicts the normalized uncertainties associated with each predicted AOT value for the same day. Higher the uncertainty, less reliable the prediction.

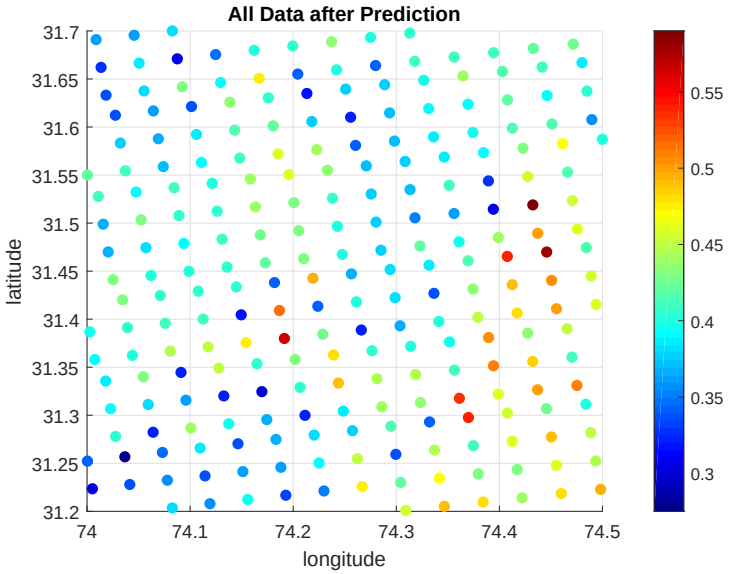

Figure 5-11: Predicting missing AOT values using the learned statistical model for $\mathcal{D}^{19}$, year 2017.

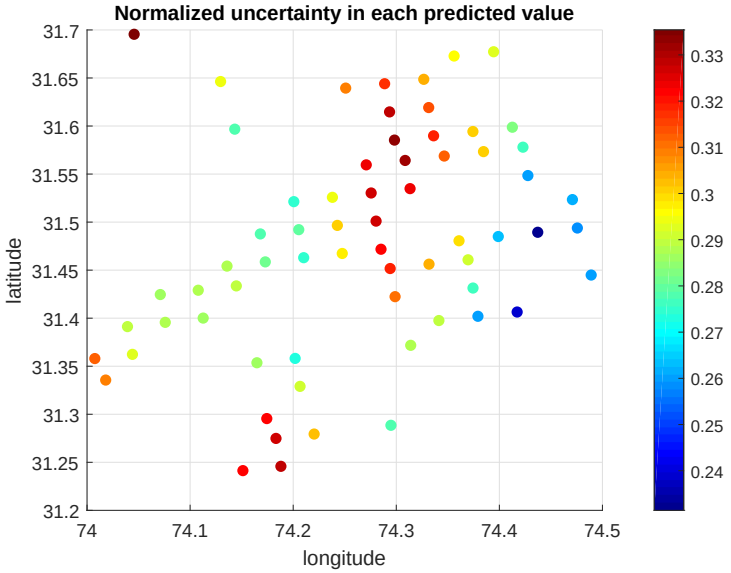

Figure 5-12: Normalized uncertainty associated with each predicted value of AOT in $\mathcal{D}^{19}$, year 2017.

In summary, it can be concluded that the problem of missing values in the

MODIS AOT data can be potentially solved using the proposed statistical model based on Gaussian Process.

5.5 Characterization of Aerosol Hot-spots

In this section, experiments related to the second research direction involving characterization of Lahore's aerosol hot-spots will be discussed. As described in the earlier chapters, the first step in achieving this objective is to identify the locations of aerosol hot-spots and then to quantify each of the hot-spots' aerosol content to obtain the complete characterization of the of the whole region. Localization, based on the Glowworm Swarm Optimization (GSO) algorithm introduced in Chapter 4, will be discussed first followed by the quantification methods.

This experiment was performed using the same dataset of MODIS AOT denoted by $\mathcal{D}^d$ for the years 2017 and 2018 belonging to the region $\mathcal{R}_{\mathcal{L}}$ of Lahore and its outskirts. Identification of hot-spot locations based on GSO will be presented next.

5.5.1 Localization of Aerosol Hot-spots

For the purpose of identifying the locations in and around Lahore that persistently show high values of AOT over time, principles governing GSO algorithm were used. Recall that GSO algorithm is useful in finding the local maxima of a multi-modal function. MODIS AOT data for each day also possesses some local maxima i.e. regions with high AOT values, which are required to be found. These local maxima were found for each day separately and their locations were overlaid on one plot to identify the potential locations in Lahore with high aerosol concentration.

Based on the working principles developed in Section 4.2.2, the algorithm that was used for carrying out localization of aerosol hot-spots is as follows:

1: Initialization

Each AOT data point from a day's dataset $\mathcal{D}^d$, was considered a glowworm and a default Luciferin-level l_0 was assigned to each of the point. Recall that in the GSO algorithm, glowworms interact with each other and the environment, ending up converging to the local maxima giving the solution of the problem. Indeed, for the AOT data, it will be the AOT points that will move and converge to the local maxima i.e. the aerosol hot-spot locations of each day. The parameters of the algorithm were initialized with the recommended values given in the Table 4.2.2, with only ρ and l_0 taking different values as suggested in [27]. Initialized values of parameters are tabulated below.

ρ	γ	β	n_t	s	l_0
0.2	0.6	0.08	5	0.03	2

The next three steps of the algorithm were repeated for 200 iterations, which were enough to ensure that the glowworms converged to the local maxima in $\mathcal{D}^d$. Moreover, value of r_s was chosen to be equal to 0.2 based on multiple experimental trials.

2: Iterative Steps of Algorithm

As described earlier in Chapter 4, Luciferin-level of each glowworm represents the closeness of that glowworm to a local maxima. Since the objective function is the two-dimensional MODIS AOT data for a day, the glowworms at the coordinates with high AOT value will get a higher Luciferin-level based on the following principle:

$$l_i(t+1) = (1-\rho)l_i(t) + \gamma y_i(t+1), \tag{5.5}$$

where y_i is the i^{th} AOT value of from the $\mathcal{D}^d$, replacing the objective function in Equation (4.1). t is the iteration number. Next, the algorithm will follow the steps outlined in the Section 4.2.2. Essentially, the movement of AOT points towards the local solution will be governed by the Equations (4.2) and (4.3) with $x_i(t) \in \mathbb{R}^2$, representing the coordinates of i^{th} AOT data point at iteration t. Before acquiring a new luciferin value, neighborhood range will be updated based on Equation (4.4).

When the above algorithm is completed, AOT data points that were spread across the whole region $\mathcal{R}_\mathcal{L}$ form a few clusters. The centroid of these clusters happen to be the local maxima of the AOT data. Therefore, the local maxima or the locations of the aerosol hot-spots were found by calculating the coordinates of these cluster centroids. When this method of localization was repeated for all days of the two years separately and the coordinates of the identified locations were overlaid, several aerosol hot-spots were discovered around Lahore. Based on the information from geographical map of $\mathcal{R}_\mathcal{L}$, shown in the Figure 5-1, these hot-spots were mapped to geographical locations manually. The hot-spots found using GSO algorithm along with the mapping (in different colors) are shown in Figure 5-13 and 5-14 for the two years. Note that the points in black were not mapped to any aerosol hot-spot.

The regions that fall within the 11 labeled hot-spots shown in the figures are listed below:

1. Fields and industries near Manga Mandi
2. Sundar Industrial Estate and Raiwind
3. Industrial Area near Valencia and Bahria Town
4. Industries near Mandi Faizabad and Mirzapur
5. Shekhupura-Sharaqpur Road
6. Industrial area of Lahore-Sheikhupura Road
7. Kala Shah Kaaku

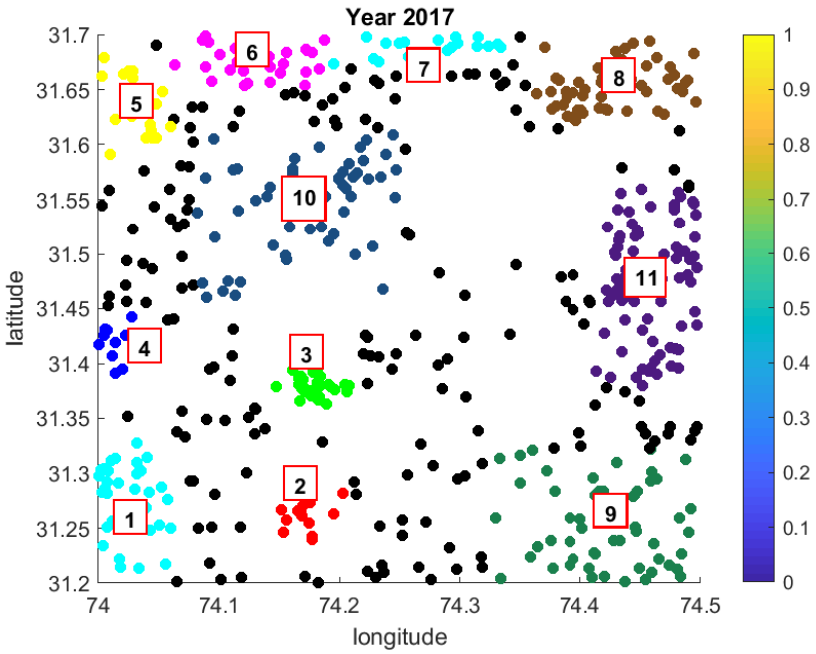

Figure 5-13: Aerosol hot-spot locations found using GSO algorithm on MODIS AOT data for year 2017.

8. Field area
9. Industrial area near Attu Asal and Mustafabad and surrounding field area
10. Lahore-Jaranwala Road (Burj Attari to Sharaqpur)
11. Field area

5.5.2 Quantification of Aerosol in Hot-spots

After having found the locations of the hot-spots, quantification of aerosol concentration within these hot-spots was also carried out based on AOT data. For the purpose of this quantification, an average value of AOT was found in the vicinity of each local maxima found. This vicinity was characterized by a radius 0.05 around each AOT point. These values are shown in Figure 5-15 and 5-16. This quantification revealed interesting results that led to a comparison of air pollu-

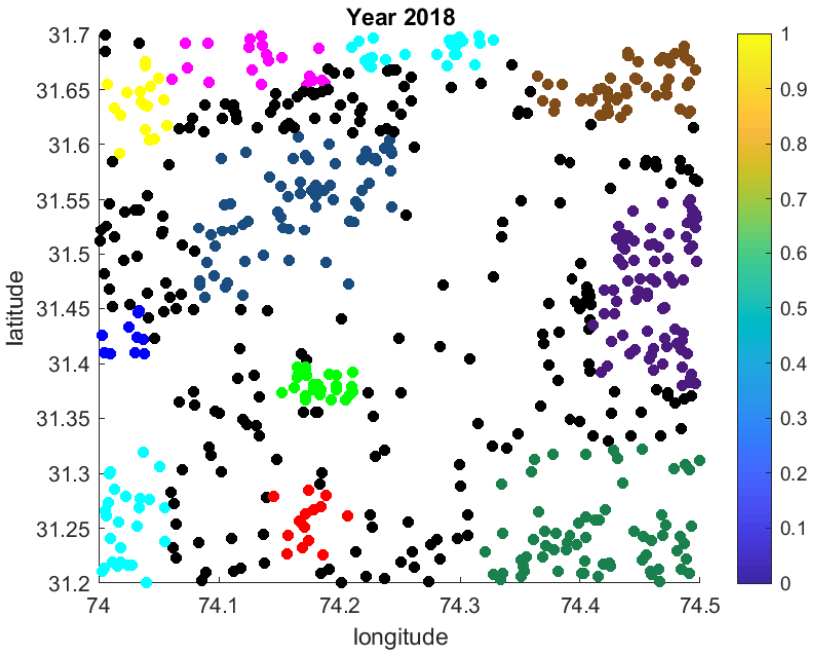

Figure 5-14: Aerosol Hot-spot locations found using GSO algorithm on MODIS AOT data for year 2018.

tion contribution made by each of these hot-spot locations. These give a fair idea about the concentration of aerosols at the local maxima that were found using GSO algorithm. To further get an estimate of aerosol content of the identified hot-spots that is easier to interpret, each of these values taken up by the points shown in the Figure 5-15 and 5-16 were averaged for each hot-spot. This led to the quantification of aerosol content of each hot-spot, the result of which is shown in Figure 5-17.

From this quantization, conclusions can be drawn about the severity of air pollution within each hot-spot. Moreover, temporal trend of aerosol concentration associated with each hot-spot can be inferred by comparing the values for the two years. For example, the aerosol concentration of hot-spot 4, the region around Mandi Faizabad and Mirzapur has a significant amount of increase in the value with time. This could reflect a new pollution source that was not present in the

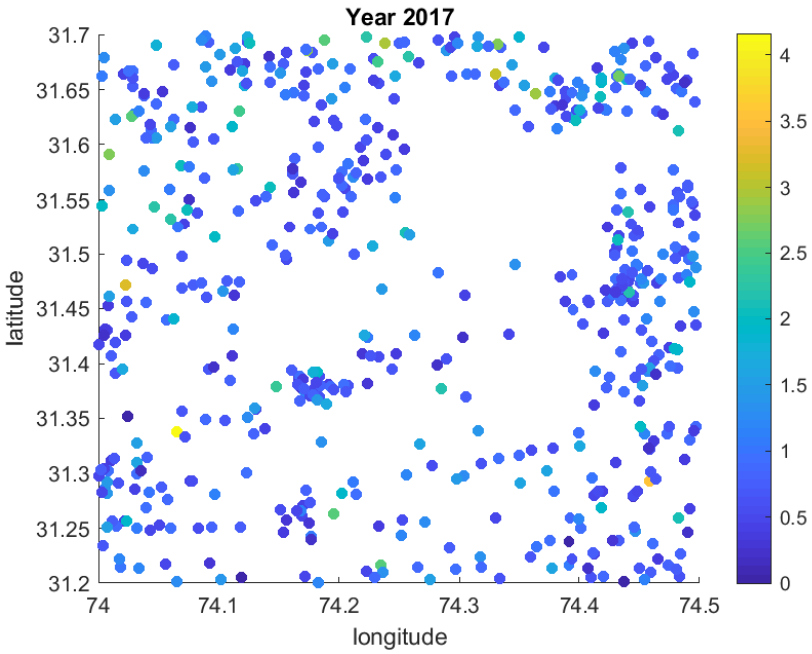

Figure 5-15: Quantification of average AOT in the vicinity of each local maxima for year 2017.

year 2017 in that region.

Discussion:

The above characterization has revealed the regions around Lahore that could be potential air pollution sources along with an estimate of each region's aerosol content. However, due to the unavailability of the ground-based sensors on these locations, it is hard to draw conclusions about the accuracy of the quantification method. The localization method seem to have worked reasonably well since the local maxima tend to map to the regions that could be potential pollution sources (e.g. industrial areas). For the purpose of evaluation of the quantification analysis, it is proposed that the relevant organizations deploy air quality sensors in these locations. Measurements from these ground-based sensors can be correlated with the estimated values to evaluate the performance of the proposed method. This will

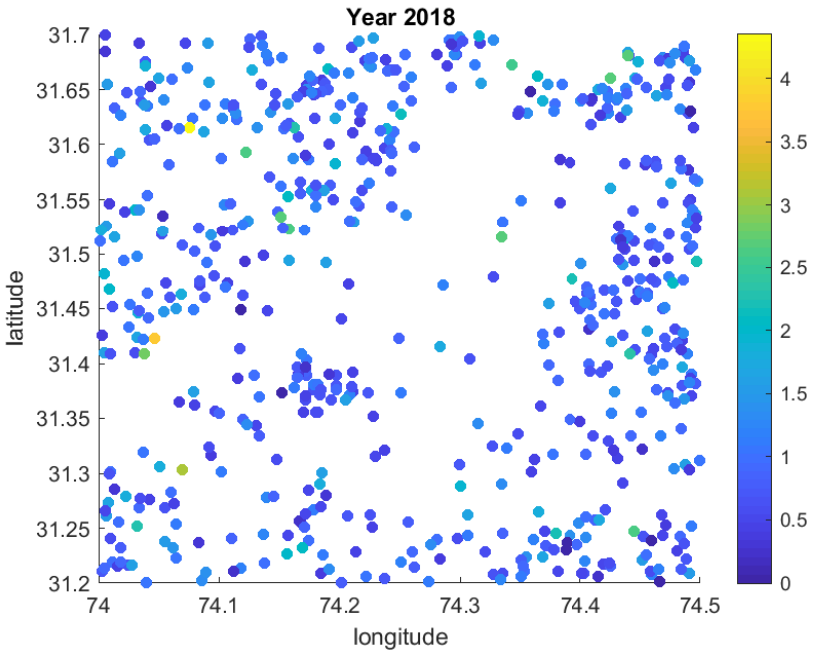

Figure 5-16: Quantification of average AOT in the vicinity of each local maxima for year 2018.

be an important step in developing a more sophisticated aerosol characterization method which is the foremost analysis to be carried out before beginning with the efforts to control the air pollution crisis.

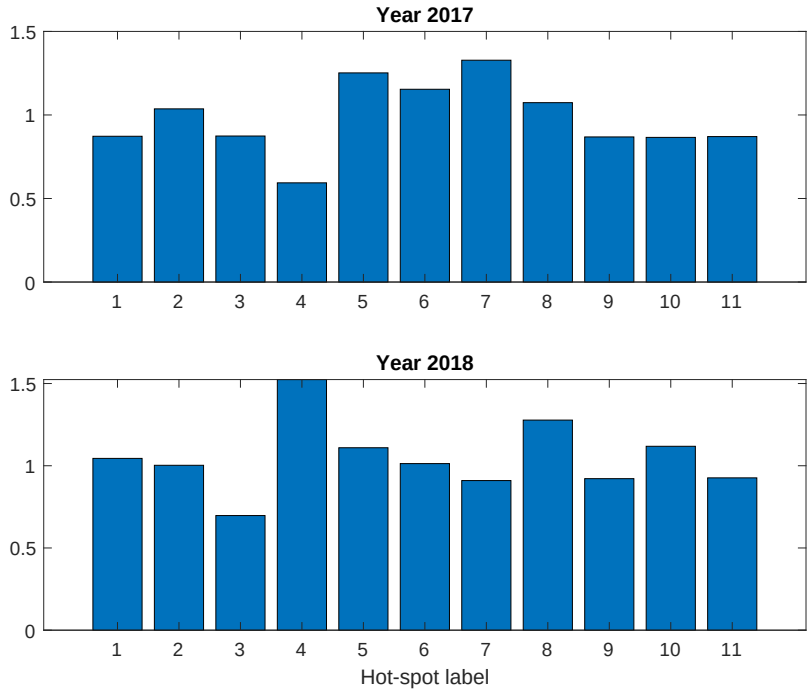

Figure 5-17: Quantification of average AOT of each hot-spot.

Chapter 6

Conclusions and Future Work

In conclusion, there were two main objectives of this work i.e., learning of a spatial statistical model for Aerosol Optical Thickness (AOT) and characterization of aerosol concentration, both for the region of Lahore and its outskirts which was carried out using satellite data of Aqua MODIS AOT 3 km data product for the years 2017-2018. To ensure the validity of MODIS data for the purpose of conducting research, positive correlation was established between MODIS AOT and ground-based AERONET AOT data.

To achieve the first objective which resolves the problem of spatial coverage holes in MODIS AOT data, a spatial statistical model based on Gaussian Processes (GP) with an ARD Exponential kernel was proposed for each day, which when validated, showed a promising cross-validation error based on normalized MSE of 0.0093 and 0.0106 for 2017 and 2018 respectively. An attractive feature of this model is that it can predict the AOT values at any location with a certain confidence level. One limitation of this part is that one cannot find a statistical model for the days when very few data points are available. Sparsity of the data points can lead to unreasonably high values of kernel hyperparameters. As a future direction, it is proposed that a spatio-temporal statistical model should be worked

out that could potentially mitigate the limitations of the current solution. This in addition to spatial, will incorporate temporal information present in the data as well.

Aerosol hot-spot characterization for Lahore, the second objective of this work, was achieved using Glowworm Swarm Optimization (GSO) algorithm, which is a meta-heuristic algorithm to locate local maxima in a multi-modal function. Application of GSO revealed several aerosol hot-spots in the region of study. Most of these hot-spots turned out to be consisting of industrial areas around Lahore, e.g. Sundar Industrial Estate and Lahore-Sheikhupura Road belonged to two of the identified hot-spots (both of which have a good number of industries). For further characterization, aerosol content in each of the identified hot-spot was estimated using a quantification metric based on a radial distance from each local maxima's center. This analysis led to a comparison of aerosol content of the hot-spots. However, it was not feasible to draw conclusions about evaluation of the proposed quantification method due to the unavailability of ground-based sensors on the hot-spot locations, using which, one can conduct correlation analysis and gauge the performance of the proposed method. Hence, it is proposed to the relevant organizations to deploy air quality sensors in these locations so the aerosol characterization method can be improved to be made more sophisticated. This could be a potential future research direction for this part of the thesis that can provide an extremely useful analysis for air quality management in the region of study.

Bibliography

[1] EPD. [Online]. Available: https://epd.punjab.gov.pk/air_quality_reports

[2] F. M. Butt, M. I. Shahzad, S. Khalid, N. Iqbal, A. Rasheed, and G. Raza, "Comparison of aerosol optical depth products from multi-satellites over densely populated cities of pakistan," *International Letters of Natural Sciences*, vol. 69, p. 12, 2018.

[3] Q. Zafar, S. Zafar, and B. Holben, "Seasonal assessment and classification of aerosols transported to lahore using aeronet and modis deep blue retrievals," *International Journal of Climatology*, vol. 38, no. 2, pp. 1022–1040, 2018.

[4] K. Alam, S. Qureshi, and T. Blaschke, "Monitoring spatio-temporal aerosol patterns over pakistan based on modis, toms and misr satellite data and a hysplit model," *Atmospheric environment*, vol. 45, no. 27, pp. 4641–4651, 2011.

[5] A. Ashraf, N. Aziz, and S. S. Ahmed, "Spatio temporal behavior of aod over pakistan using modis data," in *2013 International Conference on Aerospace Science & Engineering (ICASE)*. IEEE, 2013, pp. 1–6.

[6] F. Sharif, K. Alam, and S. Afsar, "Spatio-temporal distribution of aerosol and cloud properties over sindh using modis satellite data and a hysplit model," *Aerosol and Air Quality Research*, vol. 15, no. 2, pp. 657–672, 2015.

[7] M. Khan, B. Ghauri, and M. Bilal, "Validation of modis and misr based satellite aot data with in-situ data for lahore, pakistan," *Pakistan Journal of Meteorology*, vol. 11, no. 22, 2015.

[8] N. Hsu, M.-J. Jeong, C. Bettenhausen, A. Sayer, R. Hansell, C. Seftor, J. Huang, and S.-C. Tsay, "Enhanced deep blue aerosol retrieval algorithm: The second generation," *Journal of Geophysical Research: Atmospheres*, vol. 118, no. 16, pp. 9296–9315, 2013.

[9] R. C. Levy, L. A. Remer, D. Tanre, S. Mattoo, and Y. J. Kaufman, "Algorithm for remote sensing of tropospheric aerosol over dark targets from modis: Collections 005 and 051: Revision 2; feb 2009," *MODIS algorithm theoretical basis document*, 2009.

[10] P. Gupta, M. N. Khan, A. da Silva, and F. Patadia, "Modis aerosol optical depth observations over urban areas in pakistan: quantity and quality of the data for air quality monitoring," *Atmospheric pollution research*, vol. 4, no. 1, pp. 43–52, 2013.

[11] M. Bilal, J. E. Nichol, and M. Nazeer, "Validation of aqua-modis c051 and c006 operational aerosol products using aeronet measurements over pakistan," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 5, pp. 2074–2080, 2016.

[12] M. Ali, S. Tariq, K. Mahmood, A. Daud, A. Batool *et al.*, "A study of aerosol properties over lahore (pakistan) by using aeronet data," *Asia-Pacific Journal of Atmospheric Sciences*, vol. 50, no. 2, pp. 153–162, 2014.

[13] M. H. Shah and N. Shaheen, "Statistical analysis of atmospheric trace metals and particulate fractions in islamabad, pakistan," *Journal of hazardous materials*, vol. 147, no. 3, pp. 759–767, 2007.

[14] K. Alam, M. J. Iqbal, T. Blaschke, S. Qureshi, and G. Khan, "Monitoring spatio-temporal variations in aerosols and aerosol–cloud interactions over pakistan using modis data," *Advances in Space Research*, vol. 46, no. 9, pp. 1162–1176, 2010.

[15] M. Iftikhar, K. Alam, A. Sorooshian, W. A. Syed, S. Bibi, and H. Bibi, "Contrasting aerosol optical and radiative properties between dust and urban haze episodes in megacities of pakistan," *Atmospheric environment*, vol. 173, pp. 157–172, 2018.

[16] U. Pöschl, "Atmospheric aerosols: composition, transformation, climate and health effects," *Angewandte Chemie International Edition*, vol. 44, no. 46, pp. 7520–7540, 2005.

[17] V. Ramanathan, P. Crutzen, J. Kiehl, and D. Rosenfeld, "Aerosols, climate, and the hydrological cycle," *science*, vol. 294, no. 5549, pp. 2119–2124, 2001.

[18] J. Yang and M. Hu, "Filling the missing data gaps of daily modis aod using spatiotemporal interpolation," *Science of The Total Environment*, vol. 633, pp. 677–683, 2018.

[19] E. Sánchez-Triana, S. Enriquez, J. Afzal, A. Nakagawa, and A. S. Khan, *Cleaning Pakistan's air: policy options to address the cost of outdoor air pollution.* The World Bank, 2014.

[20] B. N. Holben, T. F. Eck, I. Slutsker, D. Tanre, J. Buis, A. Setzer, E. Vermote, J. A. Reagan, Y. Kaufman, T. Nakajima *et al.*, "Aeronet—a federated instrument network and data archive for aerosol characterization," *Remote sensing of environment*, vol. 66, no. 1, pp. 1–16, 1998.

[21] O. Dubovik and M. D. King, "A flexible inversion algorithm for retrieval of aerosol optical properties from sun and sky radiance measurements," *Journal*

of Geophysical Research: Atmospheres, vol. 105, no. D16, pp. 20 673–20 696, 2000.

[22] O. Dubovik, A. Sinyuk, T. Lapyonok, B. N. Holben, M. Mishchenko, P. Yang, T. F. Eck, H. Volten, O. Munoz, B. Veihelmann *et al.*, "Application of spheroid models to account for aerosol particle nonsphericity in remote sensing of desert dust," *Journal of Geophysical Research: Atmospheres*, vol. 111, no. D11, 2006.

[23] Terra and aqua moderate resolution imaging spectroradiometer (modis). [Online]. Available: https://ladsweb.modaps.eosdis.nasa.gov/missions-and-measurements/modis/

[24] Missions and measurements: Overview. [Online]. Available: https://ladsweb.modaps.eosdis.nasa.gov/missions-and-measurements/overview

[25] R. C. Levy, L. A. Remer, and O. Dubovik, "Global aerosol optical properties and application to moderate resolution imaging spectroradiometer aerosol retrieval over land," *Journal of Geophysical Research: Atmospheres*, vol. 112, no. D13, 2007.

[26] A. Sayer, L. Munchak, N. Hsu, R. Levy, C. Bettenhausen, and M.-J. Jeong, "Modis collection 6 aerosol products: Comparison between aqua's e-deep blue, dark target, and "merged" data sets, and usage recommendations," *Journal of Geophysical Research: Atmospheres*, vol. 119, no. 24, pp. 13–965, 2014.

[27] Y. Chen, S. Wang, W. Han, Y. Xiong, W. Wang, and L. Tong, "A new air pollution source identification method based on remotely sensed aerosol and improved glowworm swarm optimization," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 8, pp. 3454–3464, 2017.

[28] M. Nirala, "Multi-sensor data fusion of aerosol optical thickness," *International Journal of Remote Sensing*, vol. 29, no. 7, pp. 2127–2136, 2008.

[29] N. Hsu, M.-J. Jeong, C. Bettenhausen, A. Sayer, R. Hansell, C. Seftor, J. Huang, and S.-C. Tsay, "Enhanced deep blue aerosol retrieval algorithm: The second generation," *Journal of Geophysical Research: Atmospheres*, vol. 118, no. 16, pp. 9296–9315, 2013.

[30] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*. MIT Press Cambridge, MA, 2006, vol. 2, no. 3.

[31] K. Bailey. [Online]. Available: http://katbailey.github.io/post/gaussian-processes-for-dummies/

[32] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[33] M. Ebden *et al.*, "Gaussian processes for regression: A quick introduction," *The Website of Robotics Research Group in Department on Engineering Science, University of Oxford*, vol. 91, pp. 424–436, 2008.

[34] K. N. Kaipa and D. Ghose, *Glowworm swarm optimization: theory, algorithms, and applications*. Springer, 2017, vol. 698.

[35] K. Krishnanand and D. Ghose, "Glowworm swarm optimisation: a new method for optimising multi-modal functions," *International Journal of Computational Intelligence Studies*, vol. 1, no. 1, pp. 93–119, 2009.

[36] S.-C. Chu, H.-C. Huang, J. F. Roddick, and J.-S. Pan, "Overview of algorithms for swarm intelligence," in *International Conference on Computational Collective Intelligence*. Springer, 2011, pp. 28–41.

[37] [Online]. Available: https://ladsweb.modaps.eosdis.nasa.gov/search/

[38] [Online]. Available: https://aeronet.gsfc.nasa.gov/cgi-bin/webtool_inv_v3

[39] D. Duvenaud, "Automatic model construction with gaussian processes," Ph.D. dissertation, University of Cambridge, 2014.

[40] J. Quinonero-Candela, C. E. Rasmussen, and C. K. Williams, "Approximation methods for gaussian process regression," *Large-scale kernel machines*, pp. 203–224, 2007.