On remote matching of ships to pollution

Master's thesis
University of Turku
Department of Future Technologies
Computer Science
2019
Ali Leino

Supervisors:
Paavo Nevalainen
Antti Airola

TURUN YLIOPISTO
Tulevaisuuden Teknologioiden laitos

ALI LEINO: On remote matching of ships to pollution

Pro Gradu, 77 s., 4 liites.
Tietojenkäsittelytiede
Syyskuu 2019

Merenkulun päästöt ovat merkittävä haitta sekä ihmisten terveydelle että ympäristölle. Kansainvälinen merenkulkujärjestö (IMO) on asettanut rajoja käytetylle merenkulun polttoaineelle ja laivojen päästöille. Paikallaan oleva mittausasema voi mitata ohiajavien laivojen pakokaasuja ja raportoida noudattavatko laivat IMOn asetuksia. Jotta mittausasema voi toimia automaattisesti, sen pitää pystyä liittämään mitattu päästö oikeaan laivaan.
Turun saaristossa olevaa mittausasemaa käytettiin mittaamaan kaasukonsentraatioita, meteorologista dataa ja AIS-viestejä (*Automatic Identification System*) ohittavista laivoista. AIS-viesteistä käytettiin laivojen sijaintitietoja ja lähettimen luokkaa (A tai B). Vuoden 2017 dataa käytettiin mallien kouluttamiseen ja vuoden 2018 testaamiseen.
Ennustemalli kehitettiin ennustamaan laivan päästöjä vallitsevissa sääolosuhteissa. Malli perustuu Gaussin pöllähdysmalliin (engl. *Gaussian puff model*), jonka dispersioparametreja ennustetaan Pasquillin vakausluokilla (engl. *Pasquill stability class*). Jokaiselle ohi ajavalle laivalle lasketaan ennustetun konsentraation aikasarja, jonka maksimiarvo ja sen aika poimitaan. Erillistä piikintunnistusalgoritmia (engl. *peak detection algorithm*) käytettiin poimimaan vahvoja piikkejä mitatuista kaasuarvoista, joiden oletetaan tulevan laivoista. Jokainen ennustettu piikki yhdistetään ajallisesti lähimpään mitattuun piikkiin, jolloin saadaan lista laivoista ja niiden päästöistä.
Kehitettiin uusi perhe klassifiointimetriikoita, jotka ottavat huomioon mahdollisuuden sille, että yhdistämisiä tapahtuu satunnaisesti. F_1-arvosta johdettua satunnaisvirheen huomioon ottavaa F_{1t}-arvoa käytettiin metriikkana optimoitaessa yhdistämismallia. Osa mallin parametreista kontrolloi mitä mitattuja ja havaittuja piikkejä käytetään yhdistämiseen. Nämä parametrit optimointiin satunnaishaulla. Näytämme miten tulokset eivät edusta laivoja ja niiden todellisia päästöjä kun käytetään metriikkaa joka ei ota huomioon satunnaismahdollisuutta, vaan lähes mielivaltaisia yhdistyksiä. Lisäksi näytämme miten luokan B alukset huonontavat aina tuloksia.
Tuulen kalibraatiomalli suunniteltiin ja opetettiin korjaamaan lähellä olevien esteiden aiheuttamia vääristymiä mitatussa tuulen voimakkuudessa. Tuulikalibraation parametrit optimoitiin stokastisella gradienttioptimoinnilla (engl. *stochastic gradient descent*), kun satunnaishaun parametrit on ensin lukittu. Tätä prosessia vuorotellaan satunnaishaun ja gradienttihaun välillä. Tuulikalibraation tulokset validoitiin ensin visuaalisella vertailulla maastomuodostelmiin, ja sitten laskemalla neliöllinen keskiarvovirhe kalibroidun tuulen ja ilmatieteen laitoksen havaintoaseman välillä. Molemmat tulokset näyttävät, että menetelmä oppii korjaamaan tuulta niin, että se vastaa paremmin todellisia tuuliolosuhteita. Lisäksi menetelmä paransi F_{1t}-arvoa samalla vähentäen todennäköisyyttä satunnaiseen yhdistämiseen.

Asiasanat: laiva, päästö, tuulikalibraatio, dispersiomalli

UNIVERSITY OF TURKU
Department of Future Technologies

ALI LEINO: On remote matching of ships to pollution

Master's thesis, 77 p., 4 app. p.
Computer Science
September 2019

Air pollution from shipping emissions is a significant hazard for humans and the environment. International Maritime Organization (IMO) has introduced limits on the used fuel and exhaust emissions from ships. A fixed measurement station can be used to remotely measure exhaust gases from passing ships to indicate their compliance with regulations. Automatic operation of a station requires that it can pair a measurement to a ship.
Data from a measurement station in Turku archipelago was used to collect gas concentrations, atmospheric data and Automatic Identification System (AIS) measurements from passing ships. Ship location information and class (A or B) were used from the AIS data. Data for the years 2017 and 2018 was used, with the latter used as a test set.
A prediction model is designed to predict a concentration time series for a ship's route in the current atmospheric conditions. The model is based on the Gaussian puff approach with traditional Pasquill stability classes to estimate dispersion parameters. The time of maximum concentration and its value is extracted from the predictions for each passing ship. A separate peak detection model is used to extract peaks from measured gas concentrations that are significant and may originate from passing ships. Every predicted peak is matched to temporally closest observed peak to arrive at a list of possible matches of ships to their measured pollution.
A new family of classification metrics is introduced that take into account the probability of matches happening at random. A modification of the F_1-score, F_{1t}-score was used as a performance metric. One set of parameters of the model control which predicted and observed peaks are used for matching. These were optimized using Random Search (RS). We show that without the random performance adjustment the model can't be optimized to results that represent ships matched to their pollution. Further the effect of including Class B ships is shown to always reduce the performance of the model.
A novel approach is used to learn a wind calibration function to correct for the effect of nearby obstacles. With the parameters optimized by RS locked, the parameters of the wind calibration are optimized using Stochastic Gradient Descent (SGD). This process is continued with another round of RS and SGD and is hence called iterative RS+SGD. Results of the wind calibration were validated using both visual comparison and by calculating the root-mean-square error (RMSE) with a nearby weather station maintained by Finnish Meteorological Institute (FMI). The results show that the method learned a correction that better represents the true wind conditions than without the correction. Iterative RS+SGD improved the F_{1t}-score while lowering the amount of random matches.

Keywords: ship, pollution, wind calibration, dispersion modeling

Contents

List of Acronyms

AIS Automatic Identification System

CSV Comma-separated values

ECA Emission Control Area

EGCS exhaust gas cleaning system

EMGD Exponentially Modified Gaussian Distribution

FMI Finnish Meteorological Institute

FSC fuel sulphur content

GPS Global Positioning System

HFO heavy fuel oil

IMO International Maritime Organization

LNG liquid natural gas

MMSI Maritime Mobile Service Identity

$\mathbf{NO}, \mathbf{NO_2}, \mathbf{NO_x}$ Nitrogen oxide, nitrogen dioxide, and both

$\mathbf{N_2}$ Dinitrogen

REST Representational state transfer

RMSE root-mean-square error

roc ROC

RS Random Search

SGD Stochastic Gradient Descent

SO_2, SO_x Sulphur dioxide and sulphur oxides

STEAM2 Ship Traffic Emission Assessment Model

UN United Nations

WGS World Geodetic System

WMO World Meteorological Organization

List of Figures

List of Tables

Chapter 1

Introduction

Most of the global transportation of goods is done by shipping, and the reliance on shipping is still growing. Around 90% of world's merchandise by volume is transported by ships. The whole shipping industry has seen its largest growth in the last five years (2013-2018) [2].

Marine oils can have a much higher fuel sulphur content (FSC) when compared to other fuels used for transport. FSC is measured as a percentage of mass, and it can be in the order of several percentages for marine oils. For example heavy fuel oils (HFOs) are essentially a cheap waste product of oil refineries [3].

International Maritime Organization (IMO) is a specialized agency of the United Nations (UN) with 174 member states, which is responsible for setting standards in international shipping [4]. In 1997 IMO adopted MARPOL Annex VI to reduce the amount of air pollution from shipping emissions [5]. Since then they have introduced several Emission Control Areas (ECAs) which limit the maximum sulphur content of used fuel inside the areas and additionally global limits that apply outside ECAs. The global sulphur content limit has been 3.5 % since 2013, and the limit inside ECAs has been 0.1 % since 2015. In 2020, the global limit will be reduced to 0.5 % [2].

Emissions by ships are a major cause of premature deaths and childhood asthma [6]. Around 70 % of air pollution from ships are produced within 400 km of a coastline [7]. In

2007 around 60 000 premature deaths were estimated to be caused by ship emissions on global scale annually [8]. However, a recent study estimates around 400 000 premature deaths and around 14 million cases of childhood asthma are caused by the use of high-sulphur fuel annually. Even with the global low-sulphur limit taking effect in 2020, ship-based emissions are estimated to cause around 250 000 premature deaths from lung cancer and cardiovascular disease annually [6].

Ship owners have three options on how to comply with the limits set by IMO. The easiest option for old ships using traditional fuel is to use fuel with lower FSC than required. However, low-sulphur fuel is often much more expensive than high-sulphur fuel. Ships typically have multiple fuel tanks, so that they can also switch between high-sulphur fuel and low-sulphur fuel to use the cheapest alternative if they are sailing outside ECA [9].

The second alternative is to install an exhaust gas cleaning system (EGCS) [5], which can absorb around 95 % of the sulphur in the exhaust gas [10]. The requirement for ships using EGCS is that the exhaust gas may not contain more SO_x (sulphur oxides) or particulate matter than what burning compliant fuel would without the use of EGCS [11]. This allows the ship to continue using a cheaper high-sulphur fuel. However, installing an EGCS requires a substantial capital investment, from USD 5 to 9 million [11].

The third alternative which may be attractive for new ships is to use liquid natural gas (LNG), but for older ships this requires a new engine and tank setup [9]. The initial cost of converting a ship to use LNG can be from USD 6 to 22 million depending on the requirements [12].

All three alternatives have a substantial cost associated with them: either a fixed cost or an operational cost. As a consequence ship owners could have a monetary incentive not to comply with the IMO regulations if they are not monitored and enforced properly.

Enforcing the compliance of a ship can be done by the authorities by physically visiting it. This is called spot-checking. Either the process requires reading documentation proving the compliance, or by taking a fuel sample for analysis. Either way the total cost

of checking a single ship is estimated to be from EUR 300 to 400 including the cost of labour and travel, but not the cost of equipment needed for fuel analysis [13].

Checking each ship that arrives at a port is time-consuming and expensive [13]. Even if a ship has been determined to be compliant, the ship may turn from compliant to non-compliant by switching its fuel, or due to a fault in the EGCS.

An alternative to spot-checking is to use remote sensing to estimate whether a ship is compliant [14]. In remote sensing the exhaust plume of a ship is analyzed to estimate its FSC. At least two measured exhaust gases are needed to deduce the proportion of sulphur in the fuel, for example CO_2 and SO_2 (sulphur dioxide). Then the ratio $\frac{SO_2}{CO_2}$ is proportional to the FSC of the burned fuel, or in case of ships using EGCS, it is proportional to the allowed exhaust emissions [15]. This information can be used to improve spot-checking by targeting only suspected non-compliant ships determined by a sniffer result.

Sniffer methods are one way of estimating FSC by analyzing an air sample of the exhaust plume [14]. Sniffers can either be fixed installations or mobile, they only have the requirement that their gas inlet need to be physically submerged in the exhaust plume to measure it. This is why fixed sniffers should be installed near shipping lanes where ship traffic is frequent. They can be installed for example on islands, port entrances or lighthouses so that the exhaust plumes of passing ships can be sampled. Mobile sniffers have been installed at least on airplanes [16], helicopters [16], drones [13] and even vans [15]. Patrolling sniffer boats [17] have also been used, which have recently been augmented with traffic prediction to predict future traffic patterns [18].

Automatic operation of a fixed sniffer for compliance monitoring requires that it can pair a measurement to a nearby ship that is the source of the exhaust gas. Ships broadcast their positional information via radio using Automatic Identification System (AIS) messages, which are covered in more detail in section 3.1. The information broadcast by nearby ships gives rise to models that match the measurement of exhaust gas to the correct ship, which is the main objective of the model developed in this thesis.

1.1 AirNow measurement station

AirNow is an emissions monitoring service developed by KINE Robot Group [19]. It provides remote FSC measurements to authorities from fixed and mobile AirNow sniffers. Sniffers send time series data to a centralized cloud platform. The data is automatically analyzed and ship emission reports are generated from it to authorities.

This thesis uses data from one AirNow fixed sniffer station. The station contains gas sensors, meteorological sensors and an AIS receiver. Gas and meteorological measurements are averaged and stored at regular intervals of 15 seconds, called a sampling interval. In this thesis only NO_x (nitrogen oxides) are used from the gas measurements. The data is described in more detail in chapter 3.

The sniffer is located on a small uninhabited island in Turku archipelago, which is inside an ECA. Its remote location guarantees it's not close to any major pollution sources other than passing ships. The exact locations of AirNow sniffers are not public information, and are hence omitted from this thesis.

1.2 The matching problem

Air pollution modeling deals with how air pollutants from emission sources travel and disperse in the atmosphere. An air pollution model can estimate the concentration of a pollutant at a specific location, called a receptor, at specific times. These estimates can be used to generate a time series that estimates the concentration curve of the pollutant at the receptor as a function of time. Likewise the dispersion and travel of the pollution from a moving ship can be estimated. Air pollution modeling is handled in more detail in Chapter 2.

A peak is a local maximum in a time series that represents an event of interest. Using a peak detection algorithm, events of interest in a time series can be extracted. In the scope of this thesis detected peaks are suspected pollution concentrations from ships in the mea-

sured gas data. It is important to distinguish between measurements of gas concentration, which are regular readings produced by a gas analyzer, and the peaks detected from the measurements, which are more infrequent and try to represent only the measurements that originate from ships. Peak detection is handled in more detail in section 4.3.

A prediction model is introduced in Chapter 2 that estimates when a pollution from a ship route reaches its maximum concentration at a receptor. It is given ship routes and meteorological data as input, and it outputs the response curve of the pollutant at the receptor for each route separately. It may produce no estimates at all for a route, if the wind conditions are such that the pollutant can't travel to the measurement station, for example if the station is upwind of the route.

Peaks in this thesis come from two sources: predictions and observations. Sometimes a shorter form prediction or observation is used when referring to predicted peaks or predicted observations. Observed peaks are a result of a peak detection algorithm used for gas data, and can occur at any sampling step. Additionally two observed peaks can not occur at the same time. Predicted peaks are a result of prediction model applied to route and meteorological data. A single route produces either a single peak or none at all. While the concentration response curve produced by the prediction model may have multiple local maxima, for simplicity the time of maximum concentration is counted as the only peak. This may be a source of error, if a route is in a shape that naturally generates multiple maximums of the same scale (e.g. the ship passes the receptor multiple times).

Figure 1.1 shows two examples of predictions plotted with the measured concentrations. Figure 1.1 (a) shows a prediction which visually matches the measurement quite well. There is clearly one observed peak in the measurement time series. Time differences between peak maximums is only one sampling step. In figure 1.1 (b) it is unclear what the prediction should be matched to. An algorithm could detect two peaks in the measured gas values with very different magnitudes. The higher of these that reaches its maximum at 16:29 has a time difference of almost 3 minutes with the predicted peak. The

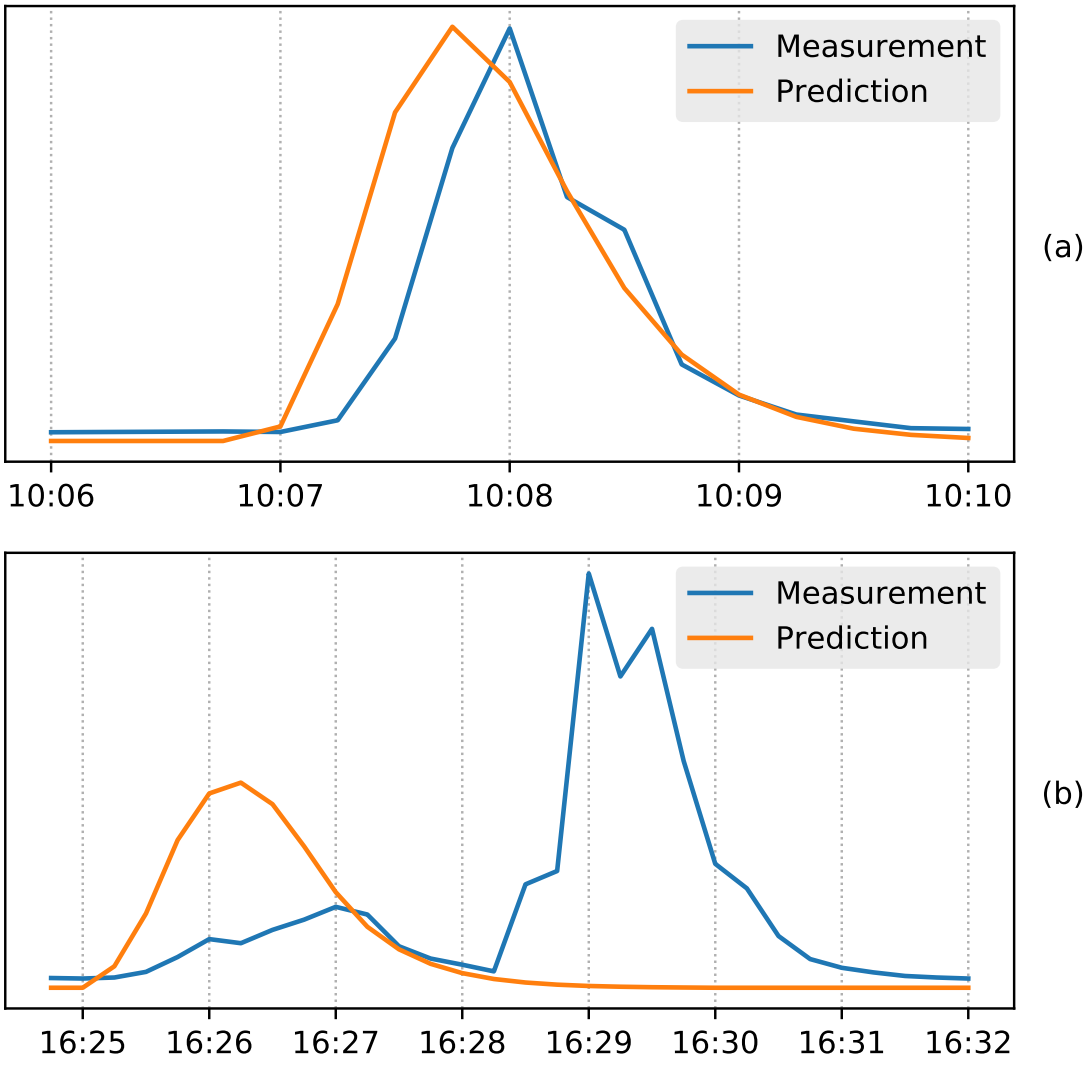

Figure 1.1: Two examples of predictions. Values have been scaled to similar levels.

smaller peak at 16:27 looks insignificant compared to it, but matches in time much better. To make matters more difficult, both of the observed peaks may be caused either by the same ship. This commonly happens when a ship sails almost parallel to the prevailing wind direction, which may result in a long period of raised concentration at the station where multiple peaks can be detected from the time series.

The goal of this thesis is to match predicted peaks from ship routes to observed peaks, so that the ship where the measurement originated from can be identified. Consequently prediction model and peak detection should produce peaks that match one another as closely as possible. The full model is illustrated in figure 1.2.

This method of matching measured air pollution to ships is indicative in nature. Both prediction model and peak detection may produce erroneous peaks. The results of both methods are but approximations of the peaks that really originated from passing ships. In this formulation of the matching problem, where only the route of the ship, meteorological conditions and gas measurements are known, the matching can never be exact. Therefore, the matching model has to be designed in a way that does not overfit to the data, and allows errors to occur.

The definition of a match has to be carefully defined to account for errors in both the predicted and observed peaks. In [20] a method is developed and tested for the ship matching problem. It is to my knowledge the only published work on the topic. They use three criteria for a successful match:

1. Both predicted and observed peaks have at least a concentration of $C_{min} = 10\,\mu\mathrm{g/m^3}$
2. The temporal difference of observed and predicted peaks is less than Δt
3. $\max\left(c_{obs}/c_{pre}, c_{pre}/c_{obs}\right) \leq C_{ratio}$ where c_{obs} observed concentration, c_{pre} is predicted concentration and C_{ratio} is the maximum ratio

The first criterion ensures both predicted and observed peaks are significant enough, and have the same limit. The second criterion ensures that the time when a peak is ob-

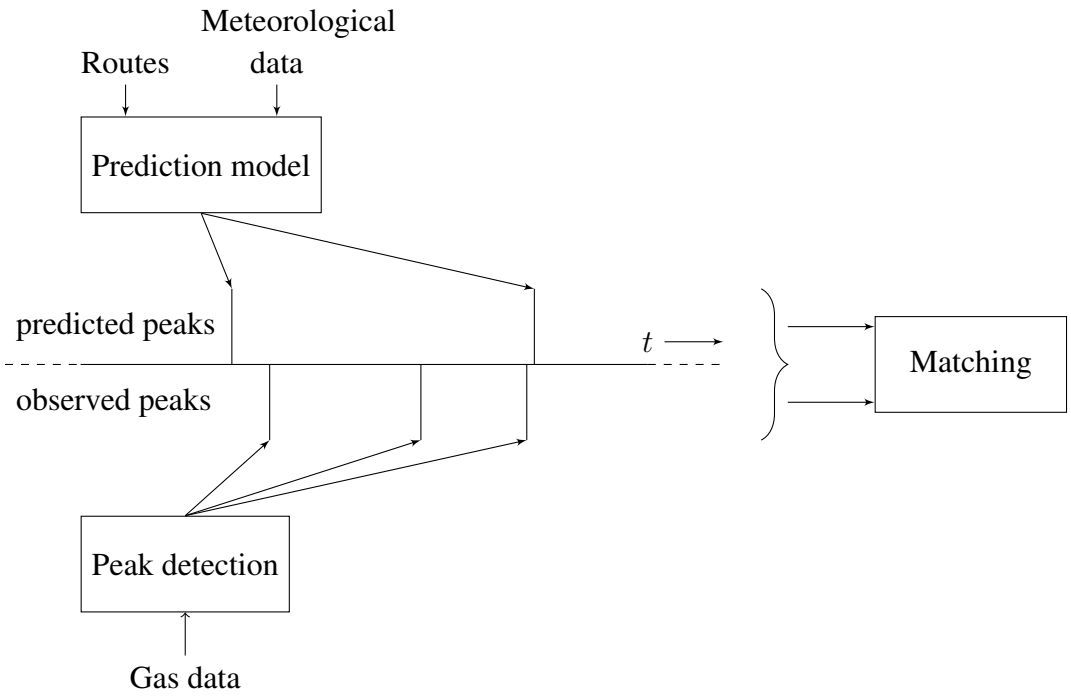

Figure 1.2: Matching model

served and predicted are not too far apart. If the time difference is large, the pollution may have come from another ship, or not a ship at all. The third criterion represents a safeguard for matches where predicted concentration differs greatly from observed concentrations. [20]

Ship Traffic Emission Assessment Model (STEAM2) can be used to estimate emission rates of ships [21], that is, the mass of produced pollutant per second a ship emits. The model relies on prior information about ships, for example their engine setup and hull characteristics [21]. The third criterion is possible because STEAM2 is used to estimate the ship's instantaneous emission rate so that predicted peaks can model observations without big differences in their relative concentrations.

In this thesis STEAM2 is not used for a few reasons. First and foremost, the third criterion can have the effect that ships that have very little prior information are not matched, because their predictions are less exact due to using default values [21]. Furthermore the error is systematic per ship, so that some ships may never be matched until their infor-

mation is updated. For the purposes of developing AirNow for compliance monitoring, this kind of biased behavior should be avoided if possible. Secondly, it is not known how good a performance can be acquired without prior knowledge about the measured ships. The limitations of not using STEAM2 or any other ship emission estimator is one of the research questions in this thesis.

In the model presented in this thesis, some of the previous matching criteria are adopted and new ones are introduced. The main difference between the model in this thesis and the model in [20] is that no limit is placed on the ratio of predicted concentrations and observations. This is because unlike in [20], there was no model used to estimate exact emission rates of ships. The matching criteria used in this thesis are:

1. Observed peak has a global significance of at least h (see 4.3)
2. Predicted peak has maximum concentration of at least p_{cmin}
3. The temporal difference of observed and predicted peaks is less than Δt

The significance of observed peaks is dependent on local and global data, and is explained in more detail in section 4.3. For predicted peaks a simple minimum concentration is used. Time criterion again ensures that predictions and observations happen close to one another in time. There is no ratio criterion that would limit matches based on concentration differences between predictions and observations.

Table 1.1 shows the basic classification of peaks according to the matching criteria. False positives occur when a peak is predicted and has a maximum concentration of at least p_{cmin}, but no observation is within time tolerance Δt. False negatives occur when a peak is observed with global significance of at least h, but no predicted peak is within time tolerance Δt. Matches occur when all matching criteria hold. These are summarized in Table 1.2. Later in this thesis, observed peaks always refer to peaks for which matching criterion 1 holds, and similarly for predicted peaks and criterion 2. Criteria 1 and 2 can be

	Peak observed	No observation
Peak predicted	True positive (a)	False positive (b)
No prediction	False negative (c)	True negative (d)

Table 1.1: Classification of observed and predicted peaks after matching .

Classification error	Criterion 1	Criterion 2	Criterion 3
True positive (a)	TRUE	TRUE	TRUE
False positive (b)	FALSE	TRUE	FALSE
False negative (c)	TRUE	FALSE	FALSE

Table 1.2: Classification errors resulting from different matching criteria being true

thought of as part of the peak detection and prediction models, and criterion 3 is the only matching criterion.

True negatives do not exist in the same sense as the other type of classifications, because the classification criteria are designed around having at least one prediction or observation with a time. True negatives are introduced later by classifying instants of time instead of peaks in continuous fashion (see Section 4.5.1). Using the error matrix in Table 1.1 simple performance metrics can be derived for the matching problem, such as recall $H = a/(a + c)$ [22].

The sets of predictions and observations can be chosen almost arbitrarily by tweaking the parameters of the matching criteria. Likewise the temporal difference for a match can be raised to a very high value to almost guarantee a match. These characteristics raise some questions about the validity of any results that the model produces. In section 4.4 these problems are discussed in more detail. A method is developed to estimate the expected number of matches that occur purely at random. These results are used to arrive at performance metrics that naturally avoid arbitrary assignment of parameters, but prefer ones which result in the least proportion of matches happening at random.

1.3 Structure of this thesis

This thesis is generally structured so that mathematical models of air pollution modeling are introduced before their applications to the matching problem. Chapter 2 gives an introduction to air pollution modeling and the relevant equations and algorithms used in the prediction model. Chapter 3 introduces the used dataset and its measurements. Exploratory data analysis is done to make decisions about the model, and to analyze the results produced later in Chapter 5. In Chapter 4 the prediction and peak detection models are described in detail, as well as the used performance metrics. Additionally, a novel approach of estimating the random probability of matches is used to adjust performance metrics with a dynamic baseline. Chapter 5 provides the results of testing the model on a test period. Chapter 6 discusses how the research questions have been answered in the thesis, and proposes some questions for further research.

Chapter 2

Air pollution modeling

Air pollution modeling deals with how air pollutants from emission sources travel and disperse in the atmosphere, and the subject has been studied extensively in the literature (e.g. [1], [23], [24]). Air pollution modeling is close to the field of atmospheric modeling, where the processes and motions of the atmosphere are modeled. The fundamental question in air pollution modeling is the movement of a pollutant gaseous material in the atmosphere. The movement can be observed by measuring the concentration of the pollutant in the air at regular time intervals at the receptor. The observed change in concentration is caused by diffusive and turbulent motions of the pollutant [25, p. 1]. Air pollution models try to estimate the concentration of a pollutant as a function of time so that it matches what is observed in nature.

While the methods in air pollution modeling are numerous with varying applications [26], this chapter only gives a brief overview of the aspects relevant for this thesis. The choices for methods and assumptions are driven by the available data at the measurement station and the requirement that the pollution sources are moving.

2.1 Gaussian dispersion equation

Gaussian dispersion equation (or Gaussian plume equation) is the most common air pollution model [23, p.561]. It calculates the concentration of a single fixed point-source emitter at a fixed receptor in stationary conditions. A collection of Gaussian-based formulas have arisen from these simple assumptions, but here only a small collection relevant to the available measurements is given. Gaussian dispersion equation can be derived as a solution to the diffusion equation in idealized conditions [27].

The gaussian dispersion equation is typically presented in a coordinate system that is aligned with the wind direction (e.g. [23], [28]). Here instead the equation is parameterized by the wind to retain geodetic alignment as was done in [1, p.141]:

$$c = \frac{Q}{\sqrt{2\pi}|\vec{u_s}|\sigma_{cw}} \exp\left(-\frac{1}{2}\left(\frac{d_{cw}}{\sigma_{cw}}\right)^2\right) C_z(z_s, z_r) \tag{2.1}$$

where:

s $= (x_s, y_s, z_s)$ is the source location,

r $= (x_r, y_r, z_r)$ is the receptor location,

c $=$ the concentration of a pollutant emitted at source s measured at receptor r,

Q $=$ source emission rate (g/s),

$\vec{u_s}$ $=$ wind speed (m/s),

d_{cw} $=$ crosswind distance from source s to receptor r (m),

σ_{cw} $=$ crosswind dispersion parameter (m),

C_z $=$ the concentration factor in the vertical direction defined later in this section

Here all equations are presented in a coordinate system with the receptor above the origin. Distances and dispersion parameters in the equation are relative to the wind direction. Additionally the source of the emission is always a ship, and the receptor is the measurement station. Because ship coordinates are mapped to a Cartesian coordinate system with the receptor centered above the origin, we have $x_r = y_r = 0$ (see Section 4.1),

which is used to simplify most equations. As usual, sea level is 0 in z-axis. Station height is 5 meters above the sea level.

Wind speed is inversely proportional to the measured concentration of a pollutant, because any clean air that mixes with a pollutant dilutes it [23]. In equation 2.1 $\vec{u_s}$ is the wind vector at the source, which may not be the same as the measured wind vector at the station $\vec{u_r}$ if the source is moving. The measured wind vector $\vec{u_r}$ is assumed to transfer the pollutant in the air, while $\vec{u_s}$ affects only the dillution of the pollution.

The crosswind distance is defined in terms of the measured receptor wind speed $\vec{u_r}$. First the downwind distance from the source to the receptor is calculated as

$$d_{dw} = \frac{\langle x_r - x_s, y_r - y_s \rangle \cdot \vec{u_r}}{|\vec{u_r}|} = -\frac{\langle x_s, y_s \rangle \cdot \vec{u}}{|\vec{u_r}|} \tag{2.2}$$

If $d_{dw} < 0$, then the receptor is $-d_{dw}$ meters upwind from the source. Negative values are useful: they can be used to determine if pollution from the source can move towards the receptor in the current wind conditions. Then crosswind distance can be calculated as

$$d_{cw} = \sqrt{|\langle x_r - x_s, y_r - y_s \rangle| - d_{dw}^2} = \sqrt{|\langle x_s, y_s \rangle|^2 - d_{dw}^2} \tag{2.3}$$

The vertical dispersion component of Equation 2.1 is

$$C(z_s, z_r) = \frac{1}{\sqrt{2\pi}\sigma_z} \left[\exp\left(-\frac{(z_s + \Delta h - z_r)^2}{2\sigma_z^2} \right) + r_g \exp\left(-\frac{(z_s + \Delta h + z_r)^2}{2\sigma_z^2} \right) \right] \tag{2.4}$$

where

σ_z = vertical dispersion parameter (m),

Δh = emission plume rise (m),

r_g = the ratio of the plume being reflected from the surface

The height of release for the pollution z_s is called the *stack height*, and the sum $H = z_s + \Delta h$ the effective stack height. The stack height for ships is not known, and is assigned a constant value of 10 meters. Plume rise is a process in which plume rises after it is

emitted. It happens as a result of the initial momentum of the exhaust gas and its higher temperature to the surrounding air [1, p.95]. It is a complex process a treatment of which would increase the prediction model complexity greatly, and is not done in this thesis. It is suspected that it has a smaller contribution to results than the stack height of the ship, which is unknown.

Equation 2.4 contains two terms of vertical dispersion. The first term is the vertical dispersion term. The first term contains an implicit assumption that the pollution can disperse vertically indefinitely, but sea level is of course very close to the source and prevents dispersion downwards very quickly. A second term is added to take into account the amount of pollution that doesn't get absorbed by the sea. The surface reflection term in a sense adds a second Gaussian distribution with the source at $(x_s, y_s, -z_s)$ below the surface [23, p. 153].

Pollution is usually assumed to be either completely reflected by the surface ($r_g = 1$) or completely absorbed ($r_g = 0$) [23, p. 562]. In this thesis complete reflection is assumed. Some formulations of the vertical dispersion term also include the reflection of the plume from the inversion layer [1].

2.1.1 Atmospheric stability

Atmospheric stability is defined as the reaction of air parcels when subject to displacements in the atmosphere. Atmosphere is said to be *stable* when an air parcel decelerates and returns to its original position when it is displaced, and *unstable* if it accelerates in the direction of displacement [26, p.52]. From these definitions it can be seen that an unstable atmosphere would result in greater dispersion of pollution, and therefore in the scope of the Gaussian equations, greater dispersion parameters.

Pasquill first proposed a discrete classification scheme in 1961 in order to define at-

[1] Inversion layer is a layer of air where the temperature increases with height, partially preventing colder air from below from mixing with it. [24, p. 191]

Table 2.1: Classification of atmospheric stability, (adapted from [1, p.148])

Stability classification	Pasquill category	Standard deviation of horizontal wind direction
Extremely unstable	A	Greater than 22.5°
Moderately unstable	B	17.5° to 22.5°
Slightly unstable	C	12.5° to 17.5°
Neutral	D	7.5° to 12.5°
Slightly stable	E	3.8° to 7.5°
Moderately stable	F	Less than 3.8°

mospheric stability based on easily available measurements from the surface [29, p.750]. It was based on measurements of wind speed, cloud cover and insolation [30]. In Table 2.1 a later modification to the classification scheme is given that is based only on wind measurements. This is because cloud cover and insolation, while more commonly used for classification, were not measured by the station. After the stability class is determined using Table 2.1, the dispersion curves for σ_h and σ_z can be determined as a function of downwind distance using the following equations, where the constants vary for each stability class [1, p.149]:

$$\sigma_h(x) = \frac{k_1 x}{(1 + x/k_2)^{k_3}} \tag{2.5}$$

$$\sigma_z(x) = \frac{k_4 x}{(1 + x/k_2)^{k_5}} \tag{2.6}$$

where x is the downwind distance in meters, σ_h is the horizontal dispersion component (both downwind and crosswind), and σ_z the vertical. For the constants k_1, k_2, k_3, k_4, k_5 the values given in [1, p.150] were used, which are based on diffusion experiments above a flat terrain.

Table 2.2: Constants for each stability class for equations 2.5 and 2.6 (adapted from [1, p.150])

Stability class	k_1	k_2	k_3	k_4	k_5
A	0.250	927	0.189	0.1020	-1.918
B	0.202	370	0.162	0.0962	-0.101
C	0.134	283	0.134	0.0722	0.102
D	0.0787	707	0.135	0.0475	0.465
E	0.0566	1070	0.137	0.0335	0.624
F	0.0370	1170	0.134	0.0220	0.700

Even though first introduced in 1961, Pasquill stability classes are still widely used [30]. The main benefit of the Pasquill approach is its original design goal: the measurements it requires are readily available [29]. Another benefit is that the constants defining stability class parameters are based on real measurements, so that the dispersion estimates are known to produce similar results in similar atmospheric conditions. This has also contributed to the success of the Gaussian models, which have been successful because of the use of dispersion parameters which can be estimated from real measurements [23, p.572].

Because of its simple nature, the Pasquill method is not robust at determining the correct atmospheric stability class at different locations. There are more precise methods for determining atmospheric stability and dispersion available [30]. However, they require either measurements from at least two nearby locations [29, p.751], or measurements that are not usually available, such as radon measurements or thermodynamic soundings [30].

2.1.2 Removal processes

Pollution can be physically removed from the air by dry and wet deposition. Dry deposition is the process in which the pollutant sticks to earth's surface and therefore no longer

moves with motions of the air [26, p.661]. Here the term dry deposition is somewhat misleading, as the surface is mostly the sea, but also the vegetation and surface in the island of the measurement station affects the dry deposition process [31, p.59]. In wet deposition the pollution is removed by absorption into droplets of water in the atmosphere [1, p.249].

Here we adopt the simple method from [1, p.172] where dry and wet deposition is modeled as a simple exponential reduction in pollutant mass by multiplying the concentration function by:

$$\exp\left(-\frac{d_{dw}}{|\vec{u_r}|\tau}\right) \tag{2.7}$$

Where $d_{dw}/|\vec{u_r}|$ is the travel time of the plume to receptor in seconds and τ is the exponential time constant. Even this very simple approximation of the removal processes is useful merely because it introduces a time component to the model, and a way to remove material from the plume and not just disperse it indefinitely.

2.2 Gaussian puff model

The Gaussian dispersion equation 2.1 assumes stationary conditions. Ships on the other hand move along a complex course, and the atmospheric conditions change with time. Ship's movement affects both the location where the pollution is released, and the dispersion of it in the air. Given the speed of the ship $\vec{v}$ as input, the observed wind speed or *true wind speed* at the ship is given by $\vec{u_s} = \vec{u_r} - \vec{v}$. As previously defined, $\vec{u_s}$ is the wind observed at the source, while $\vec{u_r}$ is the wind speed measured at the receptor. There may be differences between atmospheric conditions between source and receptor, and they can introduce some error in the model.

Puff models are based on an idea of approximating continuous emissions of air pollution into discrete packets of pollution, called puffs. The source releases these puffs at some time interval [32]. At each sampling step puffs are recalculated to allow them to

travel and transform in size and strength according to the current atmospheric conditions [33]. The concentration at the receptor at the sampling step can then be calculated using a snapshot approach, where the contributions of concentrations of the individual puffs are calculated, and their sum is the concentration at the receptor [32].

There are a few benefits in gaussian puff model compared to the continuous gaussian model. Firstly, the puff model allows atmospheric conditions to change between sampling steps easily. Secondly, it naturally extends to moving sources so that a moving source releases puffs at regular intervals [33].

Gaussian puff equation is almost identical to the Gaussian model, where only the distances are defined to the puff center:

$$C = \frac{Q}{2\pi\sigma_{dw}\sigma_{cw}} \exp\left(-\frac{1}{2}\left(\frac{d_{dw}}{\sigma_{dw}}\right)^2\right) \exp\left(-\frac{1}{2}\left(\frac{d_{cw}}{\sigma_{cw}}\right)^2\right) C_z \tag{2.8}$$

where:

C = the concentration of the pollutant at ground level (g/m^3)
Q = mass of the pollutant in the puff (g),
d_{cw} = crosswind distance from receptor to puff center (m),
d_{dw} = downwind distance from receptor to puff center (m),
σ_{cw} = crosswind dispersion parameter (m),
σ_{dw} = downwind dispersion parameter (m),
$C_z(z_p, z_r)$ = the concentration factor in the vertical direction defined in Equation 2.4,
z_p = puff center in z-direction

The definition of horizontal dispersion σ_h in Equation 2.5 assumes crosswind and downwind dispersion parameters are equal. After assuming $\sigma_{cw} = \sigma_{dw}$, the puff equation can be simplified to

$$C = \frac{Q}{2\pi\sigma_y^2} \exp\left(-\frac{1}{2}\left(\frac{d_{dw}^2 + d_{cw}^2}{\sigma_y^2}\right)\right) C_z = \tag{2.9}$$

where

σ_y = the dispersion parameter in the horizontal direction (m),

$d_{dw}^2 + d_{cw}^2$ is simply the squared distance to the puff center (m^2)

Algorithm 1 Puff dispersion model

```
1:  function PUFF-MODEL(r)
2:      puffs ← set()                          ▷ Puffs existing at current sampling step
3:      t ← min(r.t)                           ▷ Sampling step
4:      while t < max(r.t) and puffs ≠ ∅ do
5:          new-puffs ← GENERATE-PUFFS(t − Δt_s, t, r)
6:          EVOLVE-PUFFS(new-puffs, t)
7:          puffs.add(new-puffs)
8:          c_tot ← CALCULATE-MEASUREMENT(puffs)
9:          yield (t, c_tot)                   ▷ Return concentration c_tot at sampling step t
10:         REMOVE-PUFFS(puffs)
11:         t ← t + Δt_s
12:         EVOLVE-PUFFS(puffs, t)
```

Algorithm 1 shows the basic structure of a puff model algorithm for sampling a concentration time series from a route. The main loop is run at until the current sampling step t has passed the duration of the route or while puffs still exist.

At each iteration of the algorithm, new puffs in the interval $(t - \Delta t_s, t]$ are created and spawned along the route r. They may be older than the current sampling step t, so they are first evolved, so that all puffs are updated to the current sampling step. After calculating the single concentration value c_{tot} for the current sampling step, puffs that are no longer needed are removed. Finally sampling step is advanced and all puffs are evolved to the new sampling step.

Drawback of puff models is that they can be expensive to calculate [32]. A large number of simulated puffs can make the calculation of puff models infeasible [32]. In this thesis the gradient of the puff model is calculated using finite differences, which requires

the puff model to be even more performant due to the high number of evaluations of the model.

One way to limit the number of puffs is to destroy the puffs that are no longer significant for the prediction of concentration at the receptor [33]. For this thesis two removal conditions were used. The first condition is to destroy puffs that are downwind from the receptor, which can not in normal atmospheric conditions contribute any more concentration. Puffs that are more than $4\sigma_y$ upwind are removed. The second condition removes puffs that are very old. In very low wind conditions puffs may take multiple hours to reach the receptor or be triggered by the first removal rule. Puffs that are older than 60 minutes are removed.

Another strategy is to avoid calculating the contribution of a puff that is too far away to contribute significantly. As suggested in [33], puffs that are more than $3\sigma_y$ are assumed to contribute 0 concentration.

Finally, puffs should not be created at all if they are going to be destroyed in the future without any contributed concentration at the receptor. This of course can not be exactly known without running the whole puff simulation in changing atmospheric conditions. Very generous estimates were done to estimate whether a puff will ever contribute. Firstly, puffs that are more than 60 minutes away in the current wind conditions and are at least a 500 meters away are not created. The second condition is for very low meandering wind speeds, which may change quickly. Secondly, puffs that have a crosswind distance of more than 8σ from the receptor and are at least 500 meters away are not created. Again the limit is for very low wind speeds.

Algorithm 1 calculates the concentration every sampling step. Therefore, the maximum concentration time t_{cmax} also has a resolution of a sampling step. Optimization tries to minimize the differences between the predicted and observed maximum times. However, because both of these are calculated at sampling steps, it has the effect that the difference acts like a step-function, which makes the calculation of useful gradients

of the differences harder. To remedy this problem, the maximum concentration time is calculated in greater resolution than a sampling step. It is approximated by taking the neighbourhood of 3 points around the maximum time, and calculating their centre of mass with respect to time.

2.2.1 Emission rate

The emission rate of the source Q is measured in $\mathrm{g/s}$, and is an unknown function depending on multiple parameters specific for each ship. It depends on static properties of the ship such as the engine configuration, cargo load and fuel type. Some parameters affect to ship's emission rate dynamically, including acceleration of the ship, weather conditions and usage of auxillary engines. Details for estimating Q for ships are given in [10] and [21]. Because the methods require knowledge of the ship that requires an external database, they are not used in this thesis. Instead calculating Q is eliminated altogether and assigned a constant value of $100\,\mathrm{g/s}$ for all ships. This change is discussed further in Chapter 6.

Chapter 3

Data understanding

3.1 Automatic Identification System (AIS)

AIS is designed for identification of ships and tracking of their movement at sea. It was developed with the purpose of reducing the risk of collisions between ships [10]. All ships of size larger than 300 gross tonnage have to be fitted with an AIS transmitter if they make international voyages. Regardless of size all passenger ships have to be fitted with an AIS transmitter. If the ship does not operate internationally and is not a passenger ship, it can operate without an AIS transceiver up to 500 gross tonnage [34].

The ships meeting these requirements have to be fitted with a Class A AIS transceiver, which reports at least the ship's identity, position, course and speed. The transmit rate is 2-10 seconds depending on the speed of the vessel, and 3 minutes when the ship is anchored [34]. There are a significant number of smaller vessels that do not meet these requirements [10]. These ships may install a Class B AIS transmitter on a voluntary basis. Class B transceivers have a transmit rate of 30 seconds if the ship is sailing faster than 2 knots, and 3 minutes otherwise. Additionally, the transmission power of Class B units is limited to 2 W [34].

AIS uses the Global Positioning System (GPS) for positional information, and is limited by its accuracy. The accuracy is typically tens of meters [21]. Positions are given in

latitude and longitude in the World Geodetic System (WGS) coordinates [34].

The range of AIS messages is limited by many factors, including the vessel's transmitter, the receiving antenna and the geography between them . The range is therefore not constant and varies by route. This can be seen in figure 3.1, which shows the distribution of the maximum distances of received AIS messages. The data has been filtered to ships that come within 1000 meters of the measurement station. Without this limit ships that stay at a harbour nearby do not necessarily show the true range of the transmitters, because they may not move far enough.

Figure 3.1 shows that Class B transceivers have a much lower range. 95 % of the ships with Class A transmitters are detected at a range more than 7078 meters, while for Class B transmitters 95 % of the ships are detected only at a distance of 1773 meters. Class A distribution has a long tail that ends at 50 km, which has been truncated from the figure. It is possible that there were ships with class B transceivers that came within 1000 meters of the station but were never detected.

For simplicity in the following sections, ships with either Class A or Class B AIS transceivers are referred to as simply Class A or Class B ships, even though technically it is a classification of the transceiver and not the ship itself.

The distinction between the Class A and B transceivers providers two sources of information relevant to pollution modeling. Firstly, any Class B vessel is smaller than 500 gross tonnage, or 300 gross tonnage if it is on international waters[1]. Because the engine power requirement increases with the gross tonnage of the ship [35], a Class B vessel is likely to have a lower power engine. Therefore, a Class B vessel is likely to have a lower measured NO_x response in the measurement station. Secondly, Class B transceivers have a lower transmit rate and transmit power. Lower transmit rate decreases prediction performance by making routes less exact. Also partly due to transmit power the range at which messages are received is lower for class B transceivers, making routes shorter. There-

[1] unless they break the law. Class B transceivers can be bought by anyone, and can be installed anywhere.

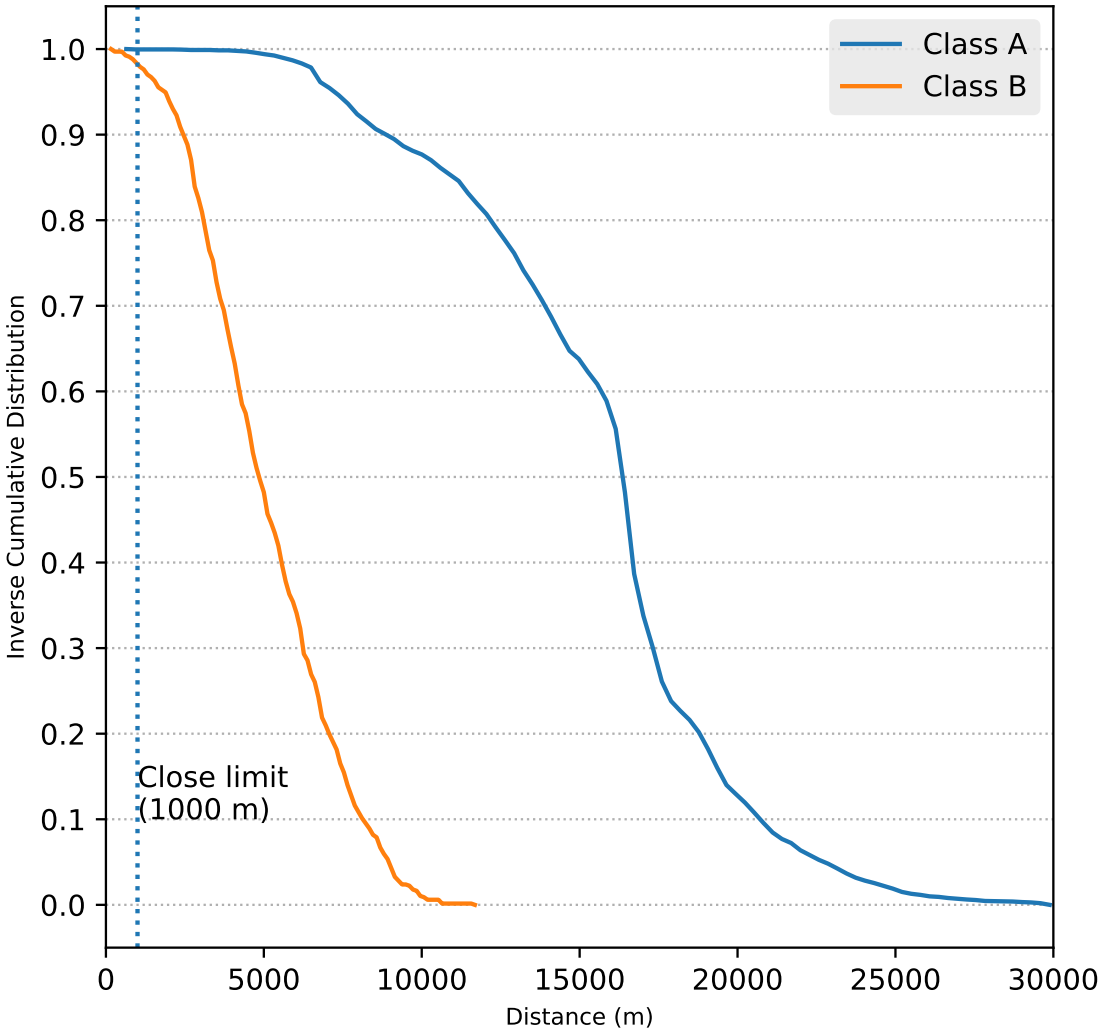

Figure 3.1: Maximum distance of received AIS messages per ship grouped by transceiver class.

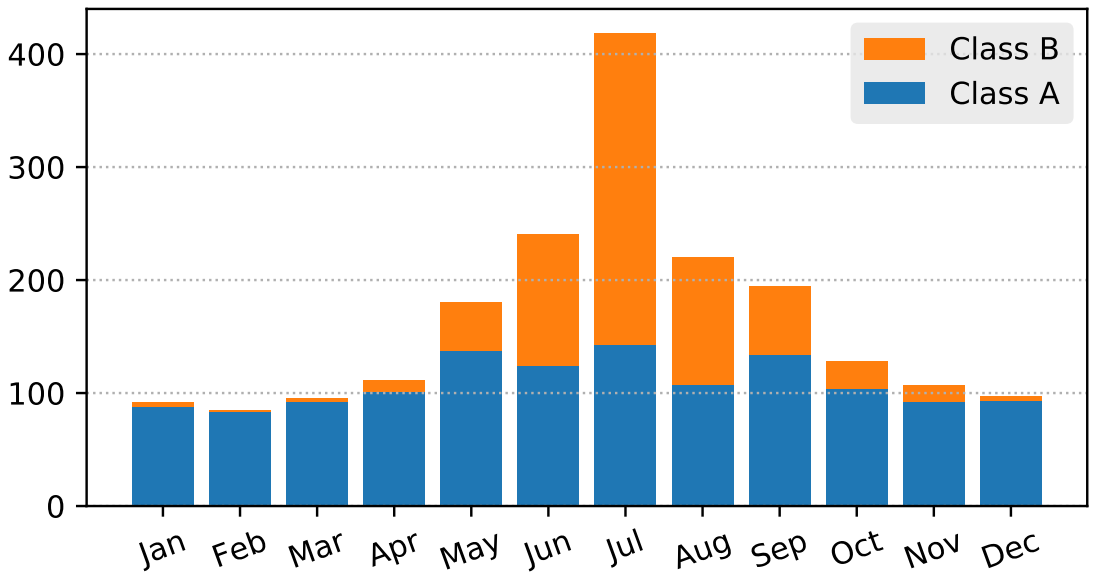

Figure 3.2: Unique monthly MMSIs recorded by the AIS receiver.

fore it is clear that the data quality of the routes depends highly on the AIS class of the transceiver.

The relative frequency of Class A and B vessels in the measurement area varies depending on the time of the year, as can be seen in Figure 3.2. The reason for this is that Class B vessels are most commonly sailing vessels and pleasure crafts (see Figure 3.4) and are unsuitable for sailing during ice cover.

Figure 3.2 also shows that the number of unique vessels is more than quadruple in July compared to February for example. While a significant number of more unique ships travel during the summer, this does not necessarily represent as big an increase in traffic volume. This can be seen in Figure 3.3, where each unique MMSI is recorded on daily basis and then the sum of the number of daily unique MMSIs is calculated for the entire month. Calculating this way, the difference between July and February is less than two-fold.

Looking only at the data in Figure 3.3 it can be expected that the model performance should be better from December to April, because there are very few Class B ships. Even

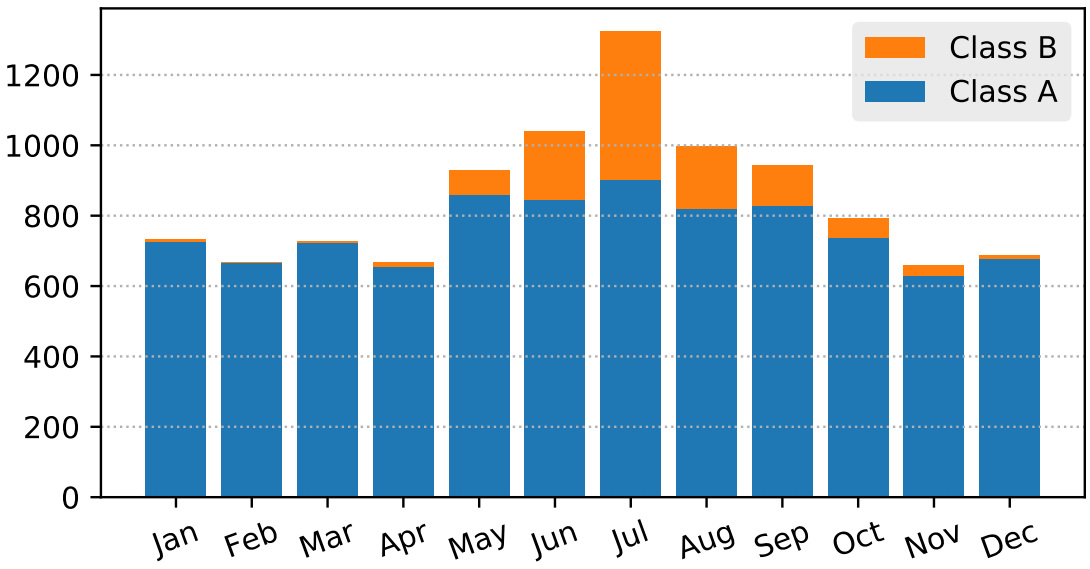

Figure 3.3: Unique daily MMSIs recorded by the AIS receiver.

if the Class B ships are filtered out, they may negatively affect the model because of their pollution, which can not be filtered out from the measured gas data.

The type of the ship is included in the AIS messages. The ship type can be used for estimating emission rate of the the ship, as was done in [21]. Yearly distribution of different ship types can be seen in Figure 3.4. Most notable is the high number of sailing vessels, which may not produce any exhaust gas if they are not running with engines.

Cumulative distribution of average vessel speeds is shown in figure 3.5. The data shows that 99.1 % of class A ships travelled at less than 20 knots. Class B vessels show a wider distribution with a significant amount of speeds around 5 to 10 knots. Around 88.9 % of class B vessels travelled less than 10 knots. Vessels with speed more than 30 knots were either search and rescue vessels, law enforcement vessels or pleasure crafts.

AIS transceivers are not required on all vessels. There is a possibility that there are some vessels whose pollution can be detected at the measurement station but are not included in the AIS data set. They are one of the limitations of an AIS-based matching model, and are considered as part of the model error. Furthermore military vessels may choose to

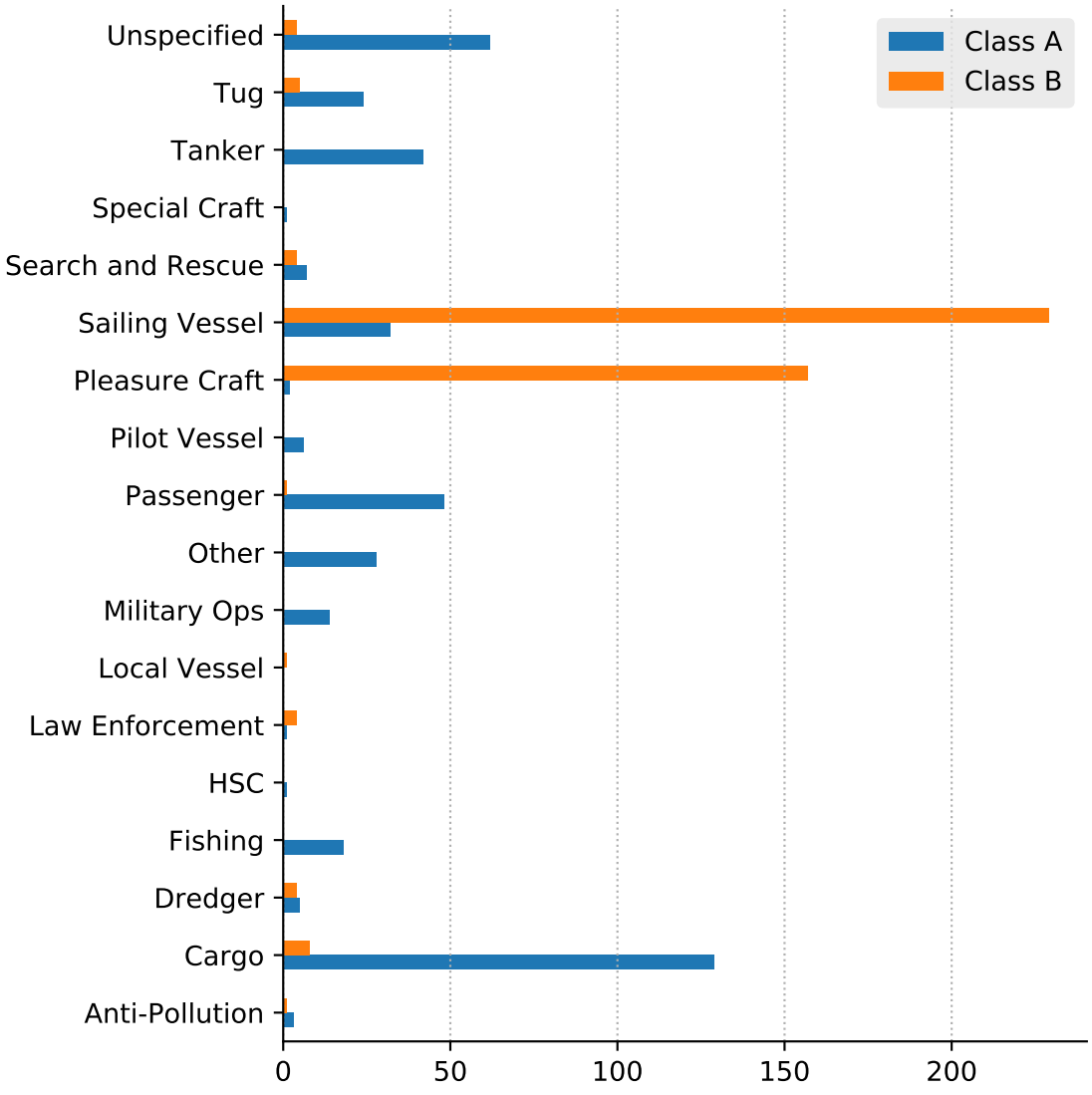

Figure 3.4: Frequency of ship types in the data set for each unique MMSI.

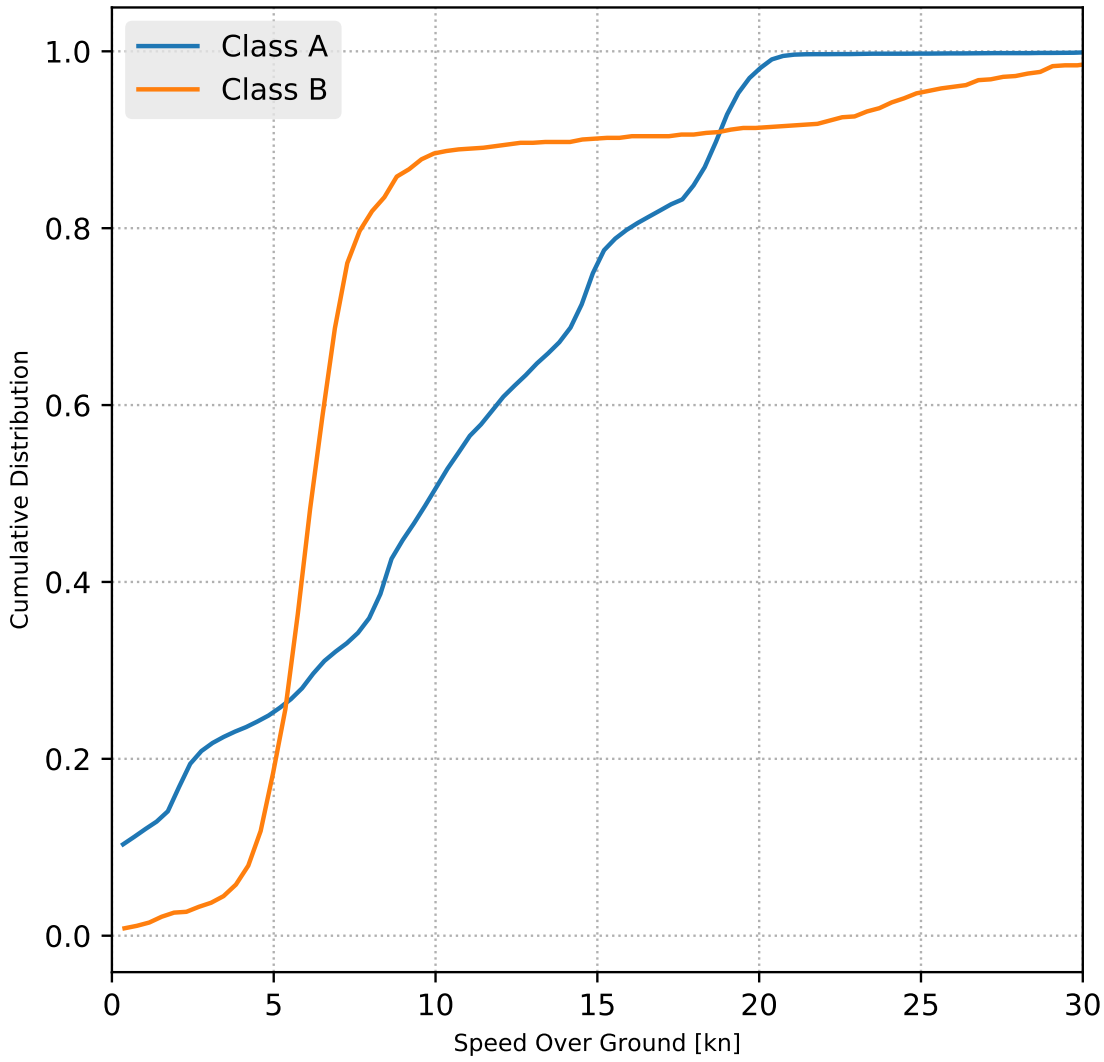

Figure 3.5: Histogram of average speeds of vessels, grouped by day and MMSI. Speeds over 30 knots have been truncated, which represent less than 0.99 % of the data.

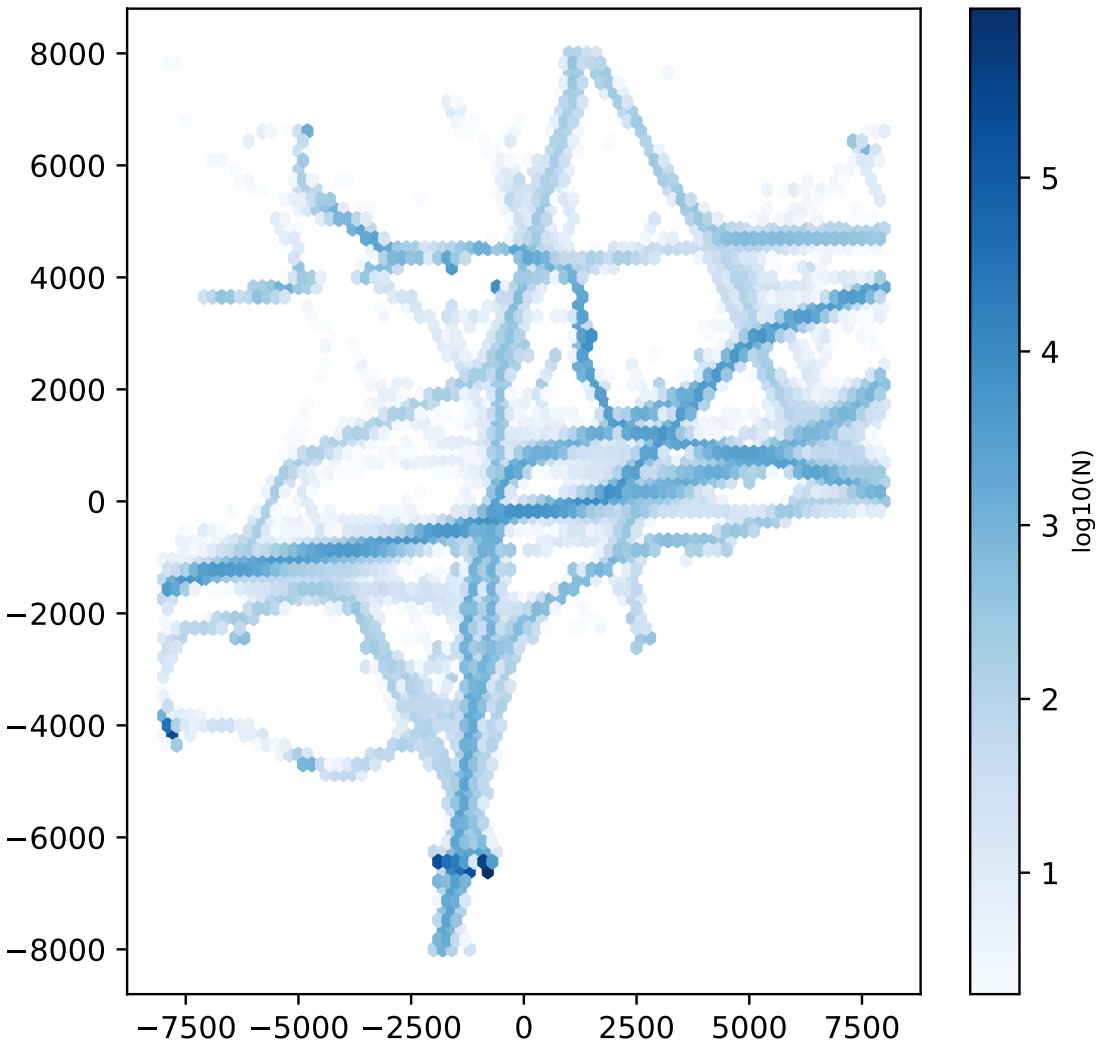

Figure 3.6: Hexagonal bin plot of AIS messages by location.

turn off their AIS transmitter altogether.

3.2 Nitrogen oxides (NO_x)

Nitrogen oxides NO and NO_2 form in the combustion of conventional fuels. Collectively nitrogen oxides are referred to as NO_x. Part of it forms as a result of high temperatures involved in combustion, which causes atmospheric N_2 (dinitrogen) to oxidise to NO and

NO_2. Another source of NO_x are compounds in the fuel that contain nitrogen [23, p. 79]. For these reasons measurements of NO_x in the atmosphere can be used to detect the combustion of conventional fuels, such as marine oils.

The measurement station contains a NO_x gas analyzer that measures the concentration of NO_x and NO separately, and calculates the concentration of NO_2 as the difference $NO_x - NO$. For the purposes of this thesis the total measurement of NO_x is used. Gas analyzer produces measurements every second, but the averaging interval (or sampling interval) is 15 seconds.

All gas sensors are checked for calibration every 24 hours for quality assurance purposes. The whole calibration process takes 30 minutes. During this time no gas measurements can be made. This means that around 2.08 % of the gas data can not be matched to any routes. During the calibration process the AIS receiver works normally, as well as meteorological sensors. The gas data during calibration is flagged as calibration data, and predicted peaks that occur during calibration are filtered (see Section 4.2.1).

3.3 Meteorological data

The measurement station has a meteorological sensor which measures air temperature, wind speed, wind direction, relative humidity and air pressure. All of these could be used in air pollution modeling [23], but for the scope of this thesis only wind speed and direction are used. The sensor produces measurements every second, which is averaged to a sampling interval of 15 seconds, which is in sync with gas measurements.

Wind direction has a strong effect on the number of routes that can result in matches. As could be seen in Figure 3.6, there is a sea lane crossing in the south-west direction of the station. If the wind blows from that direction, ships from both sea lanes can be measured by the gas analyzer. If the wind blows from north-northeast, there is only a single sea lane nearby. The further the lanes are in the current wind direction, the lower

the concentrations that are measured according to Equation 2.1.

Wind speed and direction have a strong seasonal variation . The seasonal variation during the measurement year 2017 is shown in Figure 3.7. The obstacles near the station affect the measured wind speeds and directions, which is discussed further in this thesis.

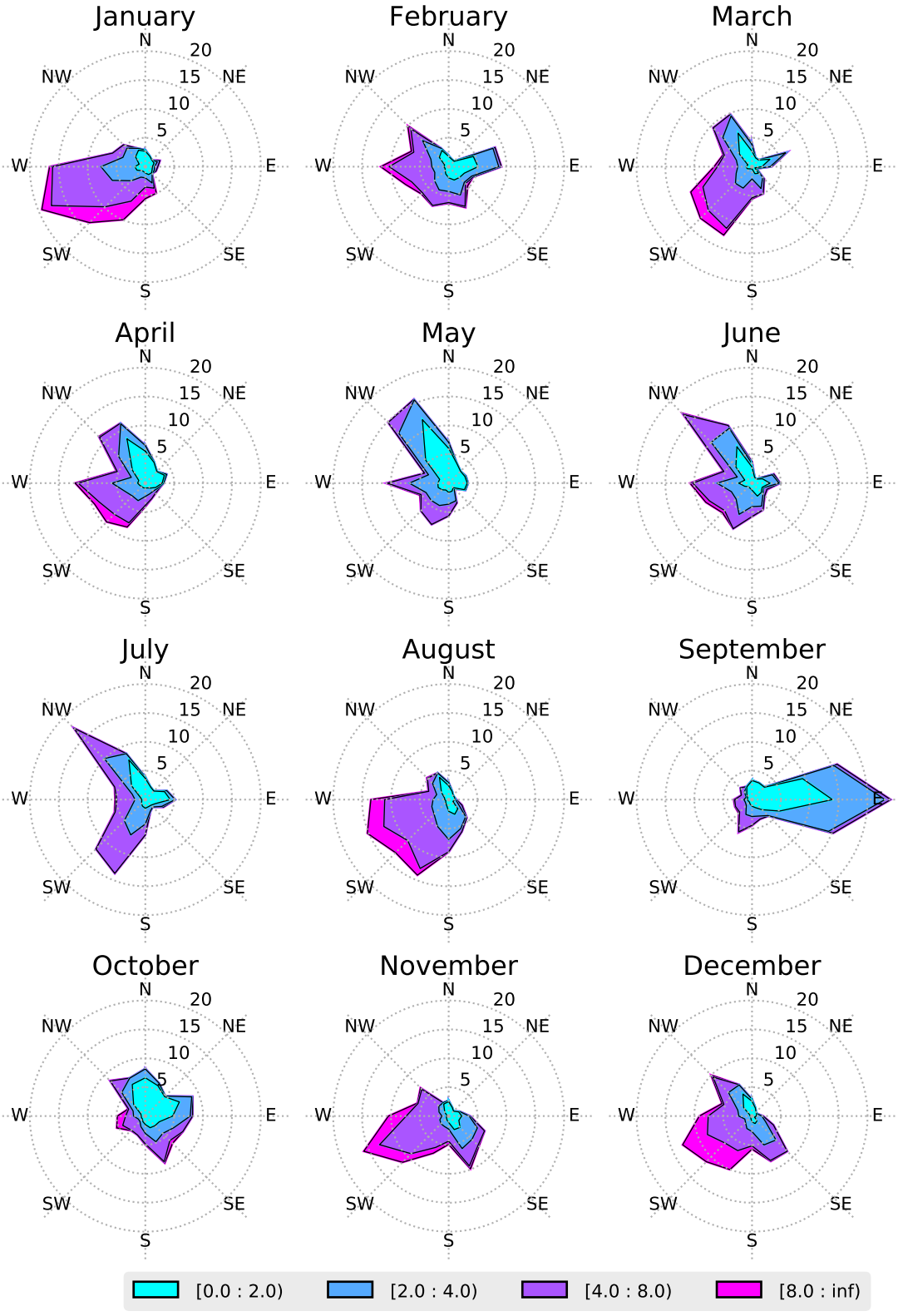

Figure 3.7: Monthly wind rose graph for the year 2017.

Chapter 4

Modeling

4.1 Coordinate system transformation

Positions in AIS messages use WGS coordinate system in latitude and longitude. The WGS coordinate system is a geodetic system, and dispersion models are typically defined in Cartesian coordinate system [28, 1, 23, 27, 24]. A was seen in Figure 3.1, the coordinates are typically less than 30 kilometers away from the measurement station. As we are limited to ships within the receiver range, there is no need for a earth-fixed coordinate system that maintains precision on the whole surface of the earth. Additionally, precision of locations kilometers away from the station is less important because the horizontal dispersion components in Gaussian model grows quickly. For example at a distance of 5 kilometers one deviation of a plume's horizontal dispersion has a radius of around 150 meters in the best Pasquill stability class, and around 1000 meters in the worst class. For these reasons a simple coordinate system transformation was adopted.

The transformation of the WGS coordinates was done to a Cartesian coordinate system using a linear projection method. Given the latitude of the station ϕ_s, longitude of the station λ_s, a constant interpolation range k and the coordinates to transform ϕ, λ, and a

WGS distance function D, the Cartesian coordinates X, Y are produced by:

$$\begin{aligned} k_\phi &= \frac{1}{k} D((\phi_s, \lambda_s), (\phi_s + k, \lambda_s)) \\ k_\lambda &= \frac{1}{k} D((\phi_s, \lambda_s), (\phi_s, \lambda_s + k)) \\ X &= k_\lambda(\lambda - \lambda_s) \\ Y &= k_\phi(\phi - \phi_s) \end{aligned} \tag{4.1}$$

The method first measures the distance from the station k degrees latitude north and divides it by k to get the approximated distance of 1 degree of latitude k_ϕ. Similarly k_λ is calculated for k degrees longitude east from the station. The values k_ϕ and k_λ are used to linearly map any latitude-pairs with the origin at the station.

Interpolation range k was chosen as 0.5, which results in k_ϕ of around 110 kilometers and k_λ of around 55 kilometers. The main goal of the coordinate transformation is to give accurate enough locations relative to the measurement station. The error of distance to the measurement station is shown in Figure 4.1. The errors have been calculated using Vincenty's formula (see [36]) as the distance function D. Vincenty's formula itself has a maximum error of about 0.5 mm [36]. The figure shows that the error is less than 8 meters at distances less than 12566 meters.

4.2 Routes

Routes model a continuous section of movement a ship makes which is recorded by the AIS receiver. The same ship may pass the station multiple times a day. Ideally a route should have a single pass of the station that can produce a single predicted peak or none at all. This is because only a single peak is extracted from a predicted concentration time series of the route.

AIS data is organized only by the time it was received. Routes are formed procedually by grouping AIS messages by MMSIs when they are read. If a new MMSI is encountered,

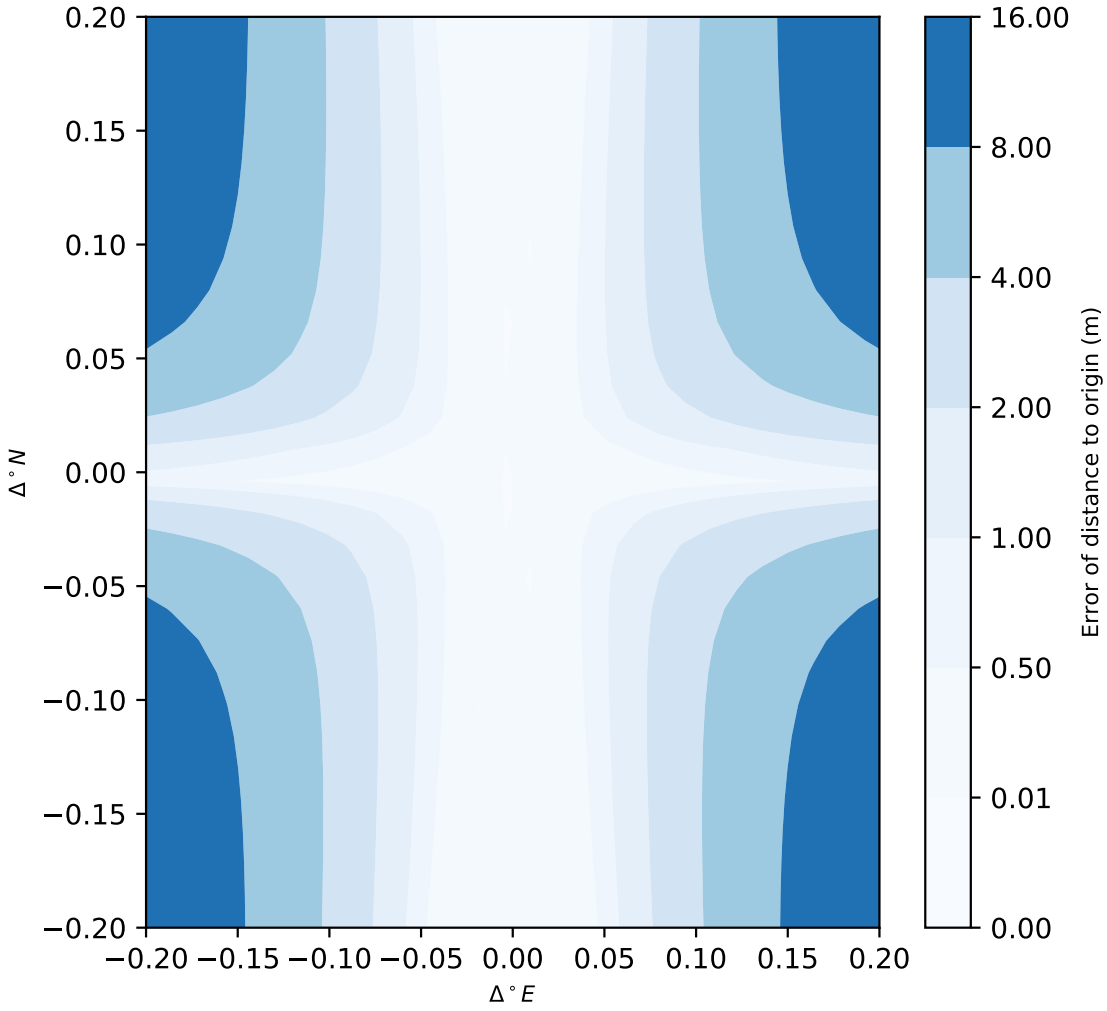

Figure 4.1: Error in distance to the measurement station for the linear coordinate system transformation of Equation 4.1 with $k = 0.5$ around the measurement station.

a new route is started and subsequent messages with the same MMSI are added to the route. Routes are marked as done when the most recently read AIS message is more than a 6 minutes from the last message in the route. 6 minutes is twice the transmit rate of a Class B vessel traveling at less than 2 knots, which still allows for a single missing value in the data (two missing subsequent missing values naturally split the route). The exact value used is 6 minutes and 5 seconds to have room for time errors.

Ships that are not moving can only negatively affect the prediction model used in this thesis. The prediction model does not take into account the speed of the vessel, even though it can possibly be mooring for extended periods of time and not produce any exhaust gas. For this reason subsequent AIS messages with no movement are removed and taken into account when splitting the route using the 5 minute interval. In other words a route can be split when a ship is not moving, even though its AIS messages are less than 5 minutes apart.

4.2.1 Route filtering

Calculating the emissions of a route is computationally expensive. The goal of route filtering is to filter out the routes that can not be matched during the optimization process. This also reduces the computational requirements of matching, because it limits the number of routes that have to be matched. Additionally parts of the route may be removed to reduce the computational time for routes. Several filters are implemented to reduce the number of routes.

First routes are filtered to ensure that routes with bad data quality are not used. Routes with less than 2 AIS measurements are not used. Increasing the lower limit risks missing Class B ships. Class B ships can be detected less than 1 km away as was seen in Figure 3.1. Furthermore ships can have have an average speed of 20 knots as was seen in Figure 3.5. At that speed during 2 AIS messages the ship would travel around 600 meters. Therefore two AIS messages may be all the station receives from a ship, even if it comes quite near

Explanation	Shorthand
Route has less than 2 AIS messages	S
Route is longer than 120 minutes	L
Route is more than 100 m upwind	U
Minimum distance to receptor is more than 8000 m	D
Maximum predicted concentration occurs during calibration	C

Table 4.1: Route filters

the station. This means that the minimum number of required AIS measurements can not be increased based on figures 3.1 and 3.5 alone, if Class B ships are part of the dataset. However, Class A ships could have a much higher minimum limit, as they are typically detected at distances of more than 10 kilometers away, and their transmission interval is 15 times faster.

Second filter limits route length to less than 120 minutes. The reason for this is that these routes are a small part of the data set, but have a comparatively huge computational time. The routes that were removed were checked manually, and they consisted of either nearby ferries which are constantly operating throughout the year, and a few sailing ships.

Equation 2.2 is used for every sampled location in the route to determine its upwind distance (negative downwind distance). All locations over 100 meters upwind are removed. If no locations remain, the route is filtered completely and is deemed filtered due to being upwind of the station.

Finally the maximum concentration time of the route is estimated using equation 2.1. If the maximum occurs during gas sensor calibration, it can not be matched to any observations, so the route is filtered to prevent a possible type c error. Puff model is not used for maximum calculation due to its slower runtime.

All the filter types are summarized in 4.1. The amount of filtered routes are summarized for the whole year of 2017 in Table 4.2. In the final version filters are applied in

Table 4.2: Filtered routes

	S	L	U	D	C	# Filtered	% of total
	✾					478	1.080
		✾				825	1.864
			✾			3179	7.182
				✾		4078	9.213
					✾	209	0.472
			✾	✾		2906	6.565
				✾	✾	90	0.203
			✾		✾	63	0.142
			✾	✾	✾	52	0.117
	✾		✾			450	1.017
	✾			✾		1065	2.406
	✾				✾	8	0.018
	✾		✾	✾		993	2.243
	✾		✾		✾	6	0.014
	✾			✾	✾	14	0.032
	✾		✾	✾	✾	19	0.043
		✾	✾			487	1.100
		✾		✾		693	1.566
		✾			✾	13	0.029
		✾	✾	✾		248	0.560
		✾	✾		✾	15	0.034
		✾		✾	✾	20	0.045
		✾	✾	✾	✾	10	0.023
						28343	64.032
Sum	3033	2311	8428	10188	519	44264	100
% of all	6.852	5.221	19.040	23.016	1.173		

order, and routes are removed when the first filter matches.

It can be seen from Table 4.2 that after all the filters are applied, around 64 % of the routes remain. The percentage of routes removed due to the calibration criterion (1.17 %) is significantly lower than the amount of calibration time per day, 2.08 %. This is by design: calibration time is hand-picked to occur when there is little traffic.

4.3 Peak detection

Peaks (or spikes) are events in a time series that are local maximum values that are significant in a particular application. The definition and characteristics of peaks are highly application dependent, and are subject to adaptation to fit a particular domain [37]. In this thesis the goal of peak detection is to detect peaks from NO_x measurements that originate from ships.

Problems in peak detection may result in both types of classification errors. If peaks are detected too often compared to the number of predictions, type c errors will increase, as there are no predictions to match them. Conversely, if peaks are detected too scarcely, type b errors will increase since some predictions are left without observed peaks. Balancing the number of peaks is not enough, as they should also represent the peaks that are caused by pollution from ships.

If we assume that a significant amount of short-term increases in concentration in the NO_x measurement time series come from ships, we can use peak detection algorithms to find a subset of the peaks correspond best to the measured ships and their predicted pollution.

Let $X = \langle x_0, x_1, \ldots, x_n \rangle$ be the sequence of values in the time series. First we define a function S which determines for a point $x_i \in X$ its "spikiness" in the subsequence $\langle x_{i-k}, x_{i-k+1}, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{i+k} \rangle$ where $k > 0$ is the size of the window. Then we calculate the spikiness function S [37]:

$$S = \frac{\max\{x_i - x_{i-1}, \ldots, x_i - x_{i-k}\} + \max\{x_i - x_{i+1}, \ldots, x_i - x_{i+k}\}}{2} \tag{4.2}$$

S calculates the maximums of the differences between x_i and each point in the window left of x_i and likewise for the window right of x_i, and then averages the maximums. This means that three points in the window contribute to the result: the center of the window x_i, the minimum value of the left window and the minimum value of the right window. This interpretation can be illustrated by writing the equation as:

$$S(x_i) = \frac{2x_i - \min\{x_{i-1}, \ldots, x_{i-k}\} - \min\{x_{i+1}, \ldots, x_{i+k}\}}{2} \tag{4.3}$$

Values of $S(x_i) < 0$ indicate that either x_i is smaller than the minimum value of the left window, or the minimum value of the right window, or both. For this reason values of $S(x_i) < 0$ are not considered as peaks. This still leaves a large amount of peaks that are not significant in either the local or the global context. For this reason two additional steps are done as in [37].

First mean μ and standard deviation σ are calculated for all $S(x_i) > 0$. Peaks are considered significant in global context if their value is at least h standard deviations above the mean, in other words the peaks for which $S(x_i) - \mu > h\sigma$.

Finally for every pair of peaks left within distance k the maximum peak is taken [37]. This discards locally insignificant peaks which are too close to another peak. Is is possible that this step removes peaks that would be useful for the purposes of analysing multiple ships whose pollution arrives in a short time window.

There are two parameters in peak detection that need to be chosen: the time window k and the number of deviations h. Time windows are restricted to from 2 to 16, which in the sampling interval of 15 seconds translates to 30 and 240 seconds. Deviations h is chosen from the interval $(0, 5]$.

4.4 Random error estimation

As a motivation for this section let's show by example how the model introduced so far can lead to undesired behaviour. Peak detection threshold h can approximately[1] control the number of observations in the range 0 to $|T|$, where $|T|$ is the number of sampling steps. Threshold h and Δt are model parameters learned from data (see 4.7). There is a possibility that the learning model tries to minimize threshold such that as few peaks are observed, and maximize Δt so that all observed peaks are matched, leading to very good apparent classification performance of the model, for example a recall of 1. This is shown to happen in practice in Chapter 5.

Therefore using simple performance metrics can easily lead to undesirable results in the model proposed in this thesis. The fundamental problem in the example was that while optimization could find a set of parameters leading to a good recall, the matches themselves were no longer representing the matching of routes to their observed pollution, but merely an assignment of routes to *any* observed pollution. The more Δt increases, the easier it is to match a prediction to an observation. This is why it can be useful to estimate the number of matches that happen purely at random. Random matches can be taken into account in the used performance metric, so that they do not count towards successful matches.

Additionally, if it the expected number of random matches is known, the results can be understood better. Even if a model with a fixed sets of observations and predictions is used, the estimation of random errors can help at understanding the quality of the results. The expected number of random matches gives a baseline performance result for the entire matching model. A novel approach taken in this thesis is that first the probability of random matches is estimated, then a performance metric is adjusted for this new baseline.

Previous work in [20] uses recall as the metric of model performance. Their model is not as prone to this problem because of a hard limit of 10 dB on the signal strength

[1]This is strictly true if there are no peaks with equal spikiness S

of observed and predicted peaks, that is, the number of observations and predictions is fixed. However, their model can still maximize Δt, which effectively maximizes random performance. In the model presented in this thesis neither the number of observations nor predictions is fixed but is allowed to be optimized by the learning model.

Random performance of a model is frequently considered in practice [38]. In ROC (roc) analysis, random performance is visualized as the baseline against which true positive and false positive rates are compared [38]. Cohen's kappa coefficient (κ) is a statistic measuring the agreement of two classifiers, which takes into account the expectation of random agreement [39]. It is the most used statistic for measuring agreement in literature [40].

4.4.1 Mathematical model

Let $|T|$ be the size of the time series of measurements (the number of sampling steps). Observed peaks are denoted by y and the set of observations as $\mathcal{Y}$. An observed peak y occurs at time $y^t \in [0, |T| - 1]$. Subscripts are used to refer to multiple observations. Observations can not occur at the same sampling step, since only one observation can occur at each sampling step.

Predictions are denoted by x and the set of predictions as $\mathcal{X}$. A prediction model should ideally make predictions in exactly the same time range as the observed peaks. Prediction model has data from the same interval as the observations, but this does not guarantee the output time range. As predictions are made to the near future, predictions can not happen before $t = 0$. Therefore, predictions are made in the range $x^t \in [0, |T| - 1 + r$, where $r \in \mathbb{Z}_{\leq 0}$. Ideally predictions should not happen more than Δt after any observations can be made, as they can never be matched. For simplicity we assume that $|T| \gg r$ so that the upper boundary is not a significant source of error and set $r = 0$ so that predictions are further limited to $x^t \in [0, |T| - 1]$.

As before the time tolerance of a match is denoted by $\Delta t \in \mathbb{Z}_{\geq 0}$. Prediction x is

matched iff

$$\exists y_j \in \mathcal{Y} \quad |y_j^t - x^t| \leq \Delta t \tag{4.4}$$

Otherwise x is not matched. Equation 4.4 allows multiple observation to match to the same prediction. This asymmetric behavior is justifiable because pollution from multiple routes can mix together and only cause a single observation, while routes and the results of the prediction model do not have this limitation.

In other words a prediction x is matched if there is an observation with time in the range $[x^t - \Delta t, x^t + \Delta t]$. However, the range may be outside the bounds of the time series T, so that the number of sampling steps within matching range of x can be less than $2\Delta t + 1$. For a prediction x with time $t = x^t$ the number of sampling steps that could have an observed peak is given by the function:

$$S_{\Delta t}(t) = \begin{cases} \Delta t + t & \text{if } |t| \leq \Delta t \\ 2\Delta t + 1 & \text{if } \Delta t < t < |T| - \Delta t \\ |T| - t + \Delta t + 1 & \text{if } |t - |T|| \leq \Delta t \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

Is is important to distinguish that the probability of a random match is not the same as the probability of a true match, as in a match that represents the true state of nature. Ideally the distribution of $\mathcal{X}$ is the same as the distribution of $\mathcal{Y}$. This means the probability that a true match occurs for prediction x should be close to 1 in an ideal model. Here we are interested in estimating the probability of matches not representing true matches. This is done by estimating the probability a match occuring at random and the expected number of such matches.

To arrive at an estimate on the number of random matches we assume $\mathcal{Y}$ is randomly distributed with respect to time. Note that only one observation can occur at each sampling step. For a given prediction x the number of matching observations k is estimated. There are $|\mathcal{Y}|$ observations which can get matched with the prediction, and the total num-

ber of possible matching time steps is given by $S_{\Delta t}(x^t)$. These assumptions result in a hypergeometric distribution for the number of matching observations k given the time of the prediction $t = x^t$:

$$Pr(t,k) = \frac{\binom{S_{\Delta t}(t)}{k}\binom{|T|-S_{\Delta t}(t)}{|\mathcal{Y}|-k}}{\binom{|T|}{|\mathcal{Y}|}} \tag{4.6}$$

And the probability of at least one observation matching to x is then

$$Pr(t) = 1 - Pr(t,0) = 1 - \frac{\binom{|T|-S_{\Delta t}(t)}{|\mathcal{Y}|}}{\binom{|T|}{|\mathcal{Y}|}} \tag{4.7}$$

The probability given by Equation 4.7 depends on both Δt and the number of observations $|\mathcal{Y}|$. $Pr(t) = 1$ when $|\mathcal{Y}| > |T| - S_{\Delta t}(t)$.

Using equations 4.5 and 4.7 the expected number of predictions with at least one randomly matching observation can be calculated:

$$\begin{aligned} E(Pr(\mathcal{X}^t)) &= \sum_{t \in \mathcal{X}^t} Pr(t) \\ &= \sum_{t \in \mathcal{X}^t} \left[1 - \frac{\binom{|T|-S_{\Delta t}(t)}{|\mathcal{Y}|}}{\binom{|T|}{|\mathcal{Y}|}} \right] \end{aligned} \tag{4.8}$$

Equation 4.8 requires the evalution of multiple combinatorials, which can be computationally expensive. However, $S_{\Delta t}(t)$ is constant in the range $\Delta t < t < |T| - \Delta t$, therefore requiring only one evaluation within that range. To guarantee the range has at least one time step, we assume that $\Delta t < |T|/2$. Then Equation 4.8 can be approximated by:

$$
\begin{aligned}
E(Pr(\mathcal{X}^t)) &= \sum_{t \in \mathcal{X}^t} Pr(t) \\
&= \sum_{|t| \leq \Delta t} Pr(t) + \sum_{\Delta t < t < |T| - \Delta t} Pr(t) + \sum_{|t - |T|| \leq \Delta t} Pr(t) \\
&\leq \sum_{|t| \leq \Delta t} Pr(\Delta t + 1) + \sum_{\Delta t < t < |T| - \Delta t} Pr(t) + \sum_{|t - |T|| \leq \Delta t} Pr(|T| - \Delta t - 1) \\
&= |\mathcal{X}| Pr(\Delta t + 1) = |\mathcal{X}| \left(1 - \frac{\binom{|T| - 2\Delta t - 1}{|\mathcal{Y}|}}{\binom{|T|}{|\mathcal{Y}|}} \right)
\end{aligned}
\tag{4.9}
$$

This approximation improves when Δt is small compared to $|T|$ and predictions are not concentrated on the edges (or outside) the range of T.

Computation of hypergeometric probability $Pr(t)$, even though having a constant number of factorials, still requires the calculation of very large factorials which will lead to issues with numerical stability in floating point calculation. For example, for a month of data with as sampling step of 15 seconds $|T| = 172800$, and its factorial is well beyond the range of 64-bit floating point value and will overflow. The following derivation solves the issue with numerical stability:

$$
\begin{aligned}
1 - Pr(t) &= \frac{\binom{|T| - S_{\Delta t}(t)}{|\mathcal{Y}|}}{\binom{|T|}{|\mathcal{Y}|}} \\
&= \exp\left[\ln \left(\frac{(|T| - S_{\Delta t}(t))!}{|\mathcal{Y}|!(|T| - S_{\Delta t}(t) - |\mathcal{Y}|)!} \cdot \frac{|\mathcal{Y}|!(|T| - |\mathcal{Y}|)!}{|T|!} \right) \right] \\
&= \exp[\ln((|T| - S_{\Delta t}(t))!) - \ln((|T| - S_{\Delta t}(t) - |\mathcal{Y}|)!) \\
&\qquad - \ln(|T|!) + \ln((|T| - |\mathcal{Y}|)!)] \\
&= \exp[\ln \Gamma(|T| - S_{\Delta t}(t) + 1) - \ln \Gamma(|T| - S_{\Delta t}(t) - |\mathcal{Y}| + 1) \\
&\qquad - \ln \Gamma(|T| + 1) + \ln \Gamma(|T| - |\mathcal{Y}| + 1)]
\end{aligned}
\tag{4.10}
$$

where $\ln \Gamma$ is the log gamma function. Equation 4.10 replaces the calculation of factorials with the logarithmic gamma function, which can be calculated directly without evaluating

the gamma function. In this thesis Stirling's approximation was used for the logarithmic gamma function. Combining equations 4.9 and 4.10 the expected number of random matches can be calculated in $O(1)$ time.

Let the expected number of random matches be $a_r = E(Pr(X^t))$. Then the expected number of true matches is simply $a_t = a - a_r$, where a is the measured number of matches (the result of the matching model). Any random match effectively removes at least one type b error. The expected number of true type b errors is then $b_t = b + a_r$. Type c error can only happen if all the matching predictions were discounted as being true matches. For simplicity this case is not considered, and the expected number of true type c errors is likewise $c_t = c + a_r$.

4.5 Performance measures

Standard F1-score is defined as $F_1 = \frac{2a}{2a+b+c}$. Simply replacing the variables with their true counterparts a_t, b_t, c_t yields F_1-score with the effect of random matches taken into account, which is here simply called true F_1-score $F_{1t} = \frac{2a_t}{2a_t+b_t+c_t} = \frac{2a_t}{2a+b+c}$. For practical purposes negative values of a_t are set to zero. Any other classification error can be treated in this way to derive its true counterpart. This can be thought of as adding a regularization term to the performance metric.

It is important to note that while F_{1t}-score resembles the F_1-score, its interpretation is slightly different. The baseline metric of a typical F_1-score is 0.5, which is the expected score of random guessing. However, F_{1t}-score already takes the baseline metric of random guessing into account, so the baseline is now 0. This is because matching happens in the time domain, where guesses are not binary. The baseline with respect to F_1-score is simply $\frac{a_r}{2a_r+b+c}$, which may be much lower than 0.5.

Automatic compliance monitoring systems are designed to be ship-centric, that is, they require a ship to which a gas measurement is assigned to. Measurements that can't

be assigned to a ship do not have the same value as measurements that can. Therefore type c errors may not matter as much as type b errors in practice. However, type c errors can be important during the optimization process, since they help the learning model to balance predictions against observations.

4.5.1 Continuous classification errors

Classification errors are discrete in nature, which is why their gradients can be hard to use in gradient-based optimization. However, the ship matching problem has the benefit of having a time component, which defines a match on a continuous axis. The previous definition of a match (see 4.4) can be modified so that Δt defines an interval of time that is matched around a pair of prediction and observation. Then type b errors occur for the time intervals that are less than Δt away from a prediction and more that Δt away from an observation. The case for type c errors is symmetric. This formulation of classification errors has the added benefit that type d errors can now be defined. A segment of time is type d if it is not within Δt to any prediction or observation. Type d errors are not used in this thesis however.

When combined with the calculation of maximum concentration times using the centre of mass approach given in section 2.2, very small changes in wind conditions can affect classification errors. The minimum change required to affect a continuous metric is 100 nanoseconds, the resolution of 64-bit DateTime objects in C♯.

4.6 Automatic wind sensor calibration

Guidelines by World Meteorological Organization (WMO) suggest that for good wind measurements the nearest obstacle should be at least 10 times further than the height of the wind instrument. Optimal installation height for wind instruments is 10 meters, or so that it is representative of the conditions in a range of a few kilometers. Any obstacles

nearby affect both the wind speed and wind direction measurements. [41, p.1.5-9]

Wind instrument at the station does not represent these optimal characteristics: it is only 2 meters above the ground, and there are obstacles such as rock formations and trees within 10 meters, and some trees within a few meters. Additionally the surface has elevation changes near the station, because the station is on a rock near the sea. Typically stations are installed so that they are easy to install and maintain, whereas a 10 meter mast would be costly to install and considerably harder to maintain.

These errors in wind direction and wind speed affect the results of the prediction model, and therefore the results of matching. The result of the prediction model should have errors that are explained by the errors of the wind measurements. These errors should be larger when wind blows from a direction with more obstacles, and lower when the wind blows from a direction with a clear path to the ship. Therefore, ship routes and the travel of their pollution gives indirect information about the effect of obstacles in the form of errors in the result. This raises the question: can the wind measurements be corrected by modeling the effects of obstacles by unknown parameters which are optimized using the matching error?

WMO suggests using multiplicative correction factors to estimate the correct wind speed when the installation location is not optimal. Obstacles lower the measured wind speed, while raising elevation of the ground strengthens the wind [41, p.1.5-11]. Errors caused by both of these effects are modeled using von Mises (or circular normal) distribution, which is the most commonly used circular distribution [42]. N von Mises distributions $P_j^{(\mathrm{vM})}(\theta|\mu_j, \kappa_j)$ are created with an initial mean $\mu_j = 2\pi(j-1)/N$ and initial concentration $\kappa_j = N$. Each distribution affects the correction multiplier C_u additively:

$$C_u(\theta) = 1 + \sum_{j=1}^{N} w_j P_j^{(\mathrm{vM})}(\theta|\mu_j, \kappa_j)) \tag{4.11}$$

where w_j are initially set to 0. Another source of error in wind measurements is from its installation, which is done by a human using a compass. Any error in the alignment of the

sensor is reflected in its measurements. This can be corrected by an additive correction factor, c_{wd}. Additionally, a non-directional multiplier c_{ws} is added. The reasoning for it is that when the concentration κ_j of a distribution changes, it modifies the multiplier somewhat for all directions. With c_{ws} it can exchange this "global" multiplier with local multiplier, possibly leading to better representation of obstacles. Nevertheless, it certainly does not weaken the model.

Finally the corrected wind speed and direction are:

$$|\vec{u}|_c = |\vec{u}| c_{ws} C_u(\measuredangle\vec{u} + c_{wd}) \tag{4.12}$$

$$\measuredangle\vec{u}_c = \measuredangle\vec{u} + c_{wd} \tag{4.13}$$

4.7 Learning model

There are multiple goals for the trained matching model. First, a preferred model is the best in terms of classification errors with the effect of random errors taken into account. Secondly, the predictions should match the observations as well as possible. In this model the former simply means that the time difference between an observation and a prediction is as small as possible. Finally, the results should indicate a useful calibration correction for the wind sensor and the timing of the gas data.

Optimized free parameters are as follows:

k = the size of the peak detection window in data points,

h = the minimum number of standard deviations for a peak's spikiness

p_{cmin} = the minimum concentration of the predicted peak for it to be considered non-zero (logarithmic),

o_{cmax} = similarly defined maximum concentration,

Δt = maximum time difference between an observation and a prediction for it to be considered a match,

c_{ws} = multiplier for wind speed used to correct multiplicative error in the measured wind speed,

c_{wd} = additive value for wind direction used to correct a constant error in the measured wind direction,

t_o = offset of time added to all predicted measurements,

$w_1, \ldots, w_N$ = weights for $P_j^{(\mathrm{vM})}$

$\mu_1, \ldots, \mu_N$ = means of $P_j^{(\mathrm{vM})}$

$\kappa_1, \ldots, \kappa_N$ = concentration of $P_j^{(\mathrm{vM})}$

Maximum concentration o_{cmax} is not useful in the case that the highest concentrations originate from the ships that are predicted. Typically this is the case, but in the next chapter the model is trained once using only Class B ships, which means there is a large quantity of peaks from larger Class A ships.

Parameters are divided into two categories, for which two different optimization approaches are used. The first category has parameters that directly affect the set of predictions and the set of observations. The parameters k, p_{cmin}, o_{cmax}, h clearly affect the predictions and observations available for matching. This results in performance metrics based on classification errors to be discontinuous when the predictions and observations change. Therefore, the search space is highly non-convex. Additionally, these parameters do not affect the results of the prediction model, only which predictions are used for matching.

The second category of parameters has continuous behavior on continuous classification errors when some limitations are placed. These parameters are c_{ws}, c_{wd}, t_o and parameters of the wind calibration model. These parameters also change the results of the prediction model. The parameter Δt is discussed later in this section.

This categorization is significant in another way. The first category of parameters do not directly affect the time series result of the prediction model, but only how it is subsequently used in matching, e.g. it is either used or not used for matching. Therefore, it is unnecessary to recalculate full route prediction when iterating values for these parameters. This reduces the calculation time considerably, as the Gaussian puff model is the most time consuming task in the full model. Likewise, parameters in the second category do not affect the observed peaks.

The first category of variables are optimized using Random Search (RS) for a maximum of 1000 iterations without improvement. Using essentially a brute-force algorithm is justifiable because of highly non-convex search space and by the fact that there are only 5 parameters to optimize. Using the above performance improvement, route predictions are calculated only once in the beginning leading to very fast performance.

Variables in the second category are optimized using Stochastic Gradient Descent (SGD). First observed peaks are calculated and cached, since none of the parameters in the second category affect them. Predictions are also calculated once to arrive at locked sets of predictions and observations. There is a possibility that changing wind multipliers may affect the maximum concentration of a predicted time series so that it is not considered for matching by p_{cmin} and o_{cmax} criteria. This is why p_{cmin} and o_{cmax} are applied only at the start of the optimization. This is a required limitation to arrive at smooth objective functions and should not affect results too much, as RS still has the power to affect the prediction set.

The result of the RS step is a set of parameters that define the prediction and observation sets. SGD however, does not control these and has to be given a fixed set of values

optimized by RS beforehand. Using some kind of unoptimized default values would be problematic, because there is no clear definition for a default value for any of the parameters. SGD optimization can not be run completely independently from the RS, because the parameters RS searches for control the set of predictions and observations. The opposite is not true: RS can be run independently from the SGD, because there are clear default values for all the variables SGD optimizes for.

In the full learning model all the parameters need to be optimized. Optimization can be alternated between running RS on the first category and SGD on the second category. When running RS for the first category, the parameters of the second category are locked to their last values and vice versa. RS uses discrete classification errors, while SGD uses a continuous version of the same classification error.

Learning rate of the SGD was updated in every alternating run using the update rule presented in [43]:

$$\gamma_t = \gamma_0(1 + 4\gamma_0 t)^{-1} \tag{4.14}$$

where γ_0 is the initial learning rate, and t is the iteration number.

The parameter Δt was left out of categorization until this point. It does not yield itself to the previous arguments for categorization, as it doesn't control the predictions or observations available for matching. On the other hand, it directly controls matches, whose maximization is an objective of both RS and SGD. Additionally, its gradient is well-defined when continuous time classification errors are used. Testing revealed that by assigning it to the first category the results converged better than when it was either in the second category or in both of them. One can argue that it belongs to the first category because it is part of random performance evaluation, and without it the random performance can't be minimized.

Parameters, their categorization and ranges are summarized in table 4.3. Parameters in the first category have more limited ranges, and the second category has practically no limits. Limits in the first category are important, as it reduces the search space for RS

Parameter	Minimum value	Maximum value	Type	Gradient scaling constant
k	2	16	integer	-
h	0	5	float	-
p_{cmin}	-10	0	float ($\log_{10}$)	-
o_{cmax}	-10	0	float ($\log_{10}$)	-
Δt	120	3600	float	-
c_{ws}	0	inf	float	0.5^{-1}
c_{wd}	$-\pi$	π	float	$(\pi/16)^{-1}$
t_o	$-\inf$	inf	float	120^{-1}
$w_1, \ldots, w_N$	$-\inf$	inf	float	2^{-1}
$\mu_1, \ldots, \mu_N$	$-\pi$	π	float	$(\pi/N)^{-1}$
$\kappa_1, \ldots, \kappa_N$	0	$2N$	float	2

Table 4.3: Summary of model parameters and their ranges. Horizontal line separates first and second category of parameters.

greatly. Some limits to κ_i are applied to prevent undefined values ($\kappa_i < 0$) or overfitting to a very tiny subset of wind direction (κ_i large). Gradient steps are scaled separately for each dimension to mimic the effect of data normalization. This is done with a scaling constant, that represents the prior belief of the range of the value, but does not limit the value in any way.

In total there are 5 parameters in the first category and $3+3N$ parameters in the second category. Only $N = 8$ is used in this thesis, leading to a parameter count of $5 + 27 = 32$.

4.8 Dataset

Table 4.4 shows the number of data points for years 2017 and 2018. There are similar numbers of wind and NO_x measurements, which is expected since they have the same

Data type	Count (2017)	Count (2018)
AIS	14 286 937	14 397 112
Routes total	44 264	50 194
Routes used	28 343	33 710
Wind	2 070 567	2 064 904
NO_x	2 084 315	2 074 990

Table 4.4: Dataset size in numbers.

sampling interval. Wind measurements include both wind speed and direction, as they come from the same sensor. Used routes refers to the amount of routes remaining after filtering.

Cross-validation is not used due to the temporal nature of the data: it is difficult to partition the data without introducing errors in the results. Ships and their pollution may end up in different folds of the data. It is also known that winds from some directions result in more ships passing at close upwind locations, as was seen in Chapter 3. For this reason every fold should have similar weather patterns balanced evenly. This requires first the analysis of what constitutes a ”weather pattern” that affects the model. For this reason, simple training and test sets are used. The year 2017 is used for training and 2018 for testing. Here an assumption is made that a year is a large enough time frame to roughly represent all common weather patterns equally.

Chapter 5

Results

5.1 Evaluation of performance metrics

As was argued in section 4.4, without careful consideration of random matching the model may perform poorly. It was argued that when maximizing recall the model could potentially maximize it to 1 so that the results would no longer reflect the matching of ship routes to their pollution. To illustrate that this is still an issue with more balanced classification metrics, the model was trained with RS maximizing F_1-score without random errors taken into account.

Training was done with the data from 2017 and tested with the data from 2018. The result of this can be seen in the contingency table 5.1, with an F_1-score of 0.668. However, the probability of a random match is $Pr(t) = 0.456$, and the expectation of random matches is $E(X) = 1442.434$. Although the contingency table looks like observations are fairly balanced with predictions, the results are rendered meaningless by the high Δt of 3 582 seconds in the result. The range for Δt was limited to 3 600 seconds, which limited how high $Pr(t)$ could raise.

In the next section the expectation of random matches is included in the performance metrics.

	Peak observed	No observation	Total
Peak predicted	2909	791	3700
No prediction	2100	0	2100
Total	5009	2100	4491

Table 5.1: Contingency table for maximizing F_1-score with no random performance adjustment

5.2 Ship class effect on performance

As was discussed in Chapter 3, Class A and B ships differ both in terms of data quality and frequency. RS was run for the parameters k, p_{cmin}, o_{cmax} and h, with the parameters in the group optimized by SGD locked at their default values. The parameter o_{cmax} was only applied for the run for Class B ships.

The months with significant amount of Class B vessels were used for training, spanning from May 2017 to (and including) September 2017. The same months from 2018 were used for testing. The reason for not using the whole year for training and testing is so that the results of Class A and B classifications are more comparable: Class B results for the whole year would represent 7 months of essentially no ships but a lot of observations belonging to Class A ships (see Figure 3.3). The whole year was tested separately for only Class A ships and when including both Class A and B ships.

Table 5.2 shows the results of the comparisons between ships with different AIS classes. Both show results above the baseline, that is $F_{1t} > 0$ (see section 4.5). Using only Class B vessels shows over a two-fold reduction in performance compared to using only Class A vessels. The models that have Class B vessels are always worse in terms of performance compared to the ones trained on Class A ships alone. This suggests that the same model should not be used for both Class A and Class B ships.

The probability of a random match is higher for Class B vessels, 0.052 for Class A

	Training		Test		
	a	F_{1t}-score	a	F_{1t}-score	$Pr(t)$
Class A (summer)	717	0.538	614	0.506	0.052
Class B (summer)	97	0.28	91	0.25	0.129
Both (summer)	752	0.504	665	0.478	0.054
Class A (year)	2204	0.55	2047	0.527	0.047
Both (year)	2197	0.536	2051	0.517	0.046

Table 5.2: The results of running RS when filtering routes by their AIS Class.

	p_{cmin}	o_{cmax}	k	Δt	h
Class A	-3.673	inf	8	287	2.012
Class B	-4.409	1.313	9	1 988	0.446
Both	-3.543	inf	8	270	1.782
Class A (year)	-3.645	inf	3	217	1.189
Both (year)	-3.671	inf	6	230	1.6

Table 5.3: Values of the free parameters for the results in Table 5.2.

vessels and 0.129 for Class B. The chosen value for Δt is multiple times higher, $1\,988$ s for Class B and 287 s for Class A case. This result suggests that the probability of random matches may get too large for some applications, a question which is further discussed in Chapter 6.

Similar results were obtained when true accuracy was optimized for instead of F_{1t}-score. These results as well as contingency tables for all the results can be seen in Appendix A.

Results for the Class A case are comparable to the results in [20], where recall is shown to increase quickly when Δt increases until Δt is around 3 minutes, after which the slope gets flatter. Only Class A ships are used in their model. They choose 5 minutes to achieve a recall of 0.5. They measured 132 peaks in two months. With these we can calculate that the probability of a random match is 0.016. Their lower probability of a random match can be explained by the dataset: the traffic density in the dataset of this thesis is over 3.5 times higher, 298 ships per month versus 81 ships per month in their dataset. It should be noted that exact comparisons are hard to make since their results do not include performance in a separate test set.

5.3 Alternating RS and SGD

RS and SGD were run subsequently for a total of 30 iterations. Wind calibration function $C_u(\theta)$ (see Equation 4.11) was trained with $N = 8$ von Mises distributions. SGD was run with a batch size of 0.1 for one epoch with an initial learning rate of 0.15. First iteration is RS and the last iteration is SGD. The model was trained on Class A ships from 2017 and tested with Class A ships from 2018.

Figure 5.1 (a) shows the change in the objective function F_{1t}-score per iteration. Scores in both training and test improve until about iteration 5, and then either lower or stagnate. This is likely caused by the slight difference in the values given by continu-

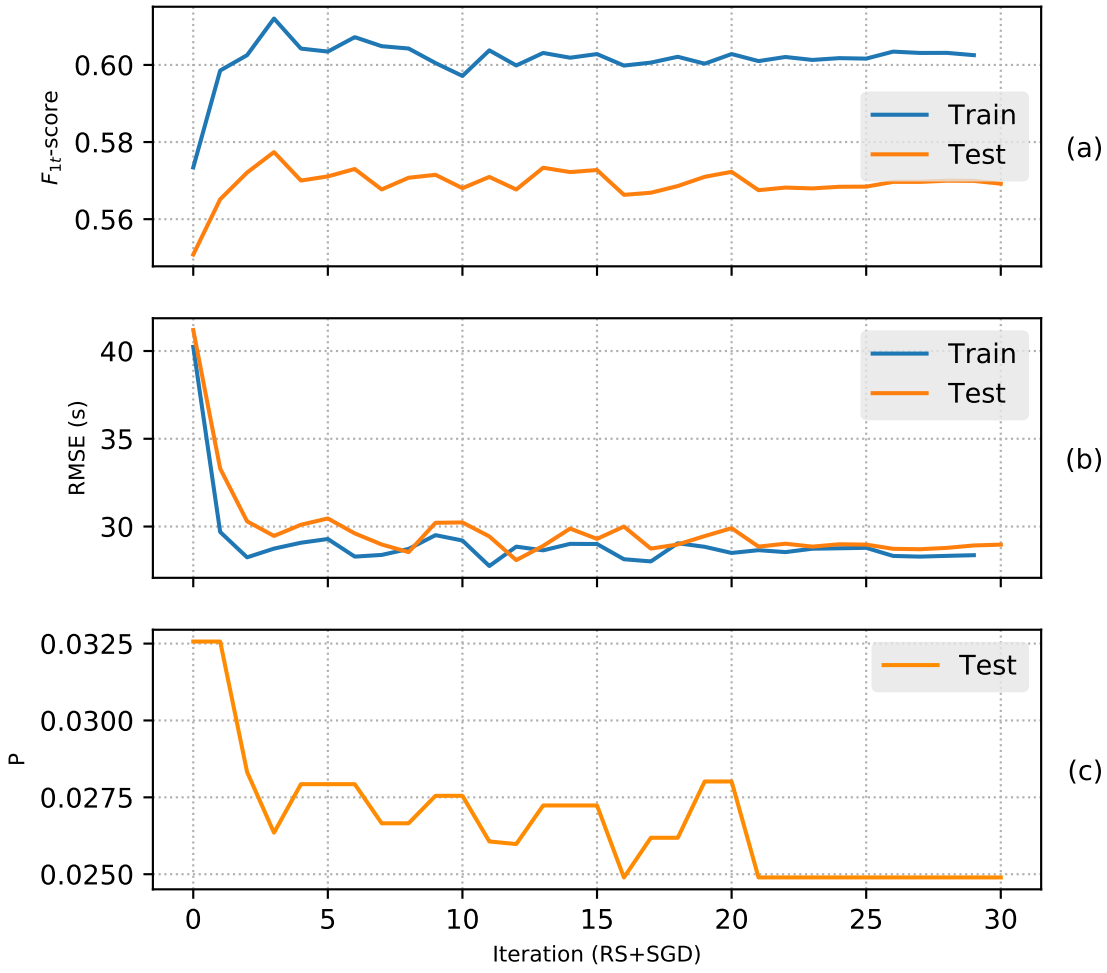

Figure 5.1: Results of running RS+SGD for 30 iterations. (a) Train and test F_{1t}-scores as a function of iteration. (b) RMSE-error of the time difference of matches between predicted and observed peaks. (c) Probability of a random match.

ous and discrete F_{1t}-scores. Figure 5.1 (b) shows the RMSE error in the timing difference in matches. It is improving slightly even after iteration 5. Finally Figure 5.1 (c) shows the positive tradeoff happening in later iterations: The probability of random matches gets lower even when the value of the objective function decreases slightly.

The final values for the test set are 0.569 F_{1t}-score, RMSE of 28.968 s, and $Pr(t) = 0.025$. These results should be compared against the previous result in Table 5.2 for Class A ships for the full year. F_{1t}-score had a relative increase of 7.97 % (0.569 versus 0.527). $Pr(t)$ had a a relative decrease of -46.81 % (0.025 versus 0.047). This improvement is the most significant of these, as it means that the matches are half as likely to occur at random.

The difference between F_{1t}-scores on training and test sets were substantial. This result is present also in previous results shown in Table 5.2. However, the generalization error does not seem to be getting worse in later iterations. The generalization error is not so pronounced in RMSE-errors, where the test set sometimes even has better performance, which suggests that the timing of predictions versus observations is generalizable, and something else causes the difference. It may be caused by the tight limits learned in the RS step, which could cause worse observed peaks from getting extracted and matched. Differences in gas measurements could cause this, if for example the noise in the gas measurements changed. Improving the generalization could be studied in future works.

Wind calibrations were learned in each of the SGD steps of the algorithm. The progress of the wind calibration function $C_u(\theta)$ can be seen in Figure 5.2. Before the first iteration the function is a unit circle with 8 von Mises distributions evenly spaced. The mean of each distribution is shown as a line. The green or red shading shows the difference between the last shown iteration. The basic shape of the function is learned mostly in one iteration, and most of the graph has converged in 7 iterations. After 25 iterations, no visual changes can be detected. Most of the multipliers of the distributions are positive. The distribution shown with red line gets a negative multiplier, and the to-

tal value of C_u to southeast is less than 1, so that it weights the wind negatively. Two approaches are taken to interpret the results.

Previous wind calibrations can be compared to the measurement station by visual inspection. A picture of the measurement station is shown in Figure 5.3. First iteration seems to mostly show the effect of the trees in northwest and east. It is surprising how large the multipliers are in the southeastern direction. It may be mostly caused by a single tree, about 2 meters from the station, whose top is slightly above the wind sensor.

The negative multiplier for the distribution shown with a red line in southwestern direction dampens measured wind speed. This can be explained by geography: there are no obstacles in that direction, and there is a raising elevation from sea level, which strengthens wind speeds [41, p.1.5-11].

Figure 5.2 also shows that high multipliers were learned for the northeastern direction. But the geography would suggest they should be substantially larger than any other multipliers, because there is a forest with trees around 10 meters tall that direction. This could be explained by the fact that its dampening effect is too high for any wind speed to register at all from that direction. This is supported by Figure 3.7, which shows northeastern winds never exceeding 2 m/s. The fact northeastern winds are so infrequent also explains why the final multipliers were learned in later iterations. WMO states that measurements made in the wake of a tree row have little information about the true wind conditions [41]. Finally, the forest may cause a wake between it and the station, and the pollution simply moves past the station without it being measured.

The end goal of improving wind measurements is not to provide accurate wind measurements in general, but wind measurements that reflect the average wind conditions that move the pollution from ship to the station. The model is not trained on true wind speeds, but on the errors on the predicted and measured peak timing error. Therefore, it may be that the forest in the northeast receives similar multipliers as southeast for a simple reason: pollution truly moves more slowly because of the forest, but a single tree hardly affects

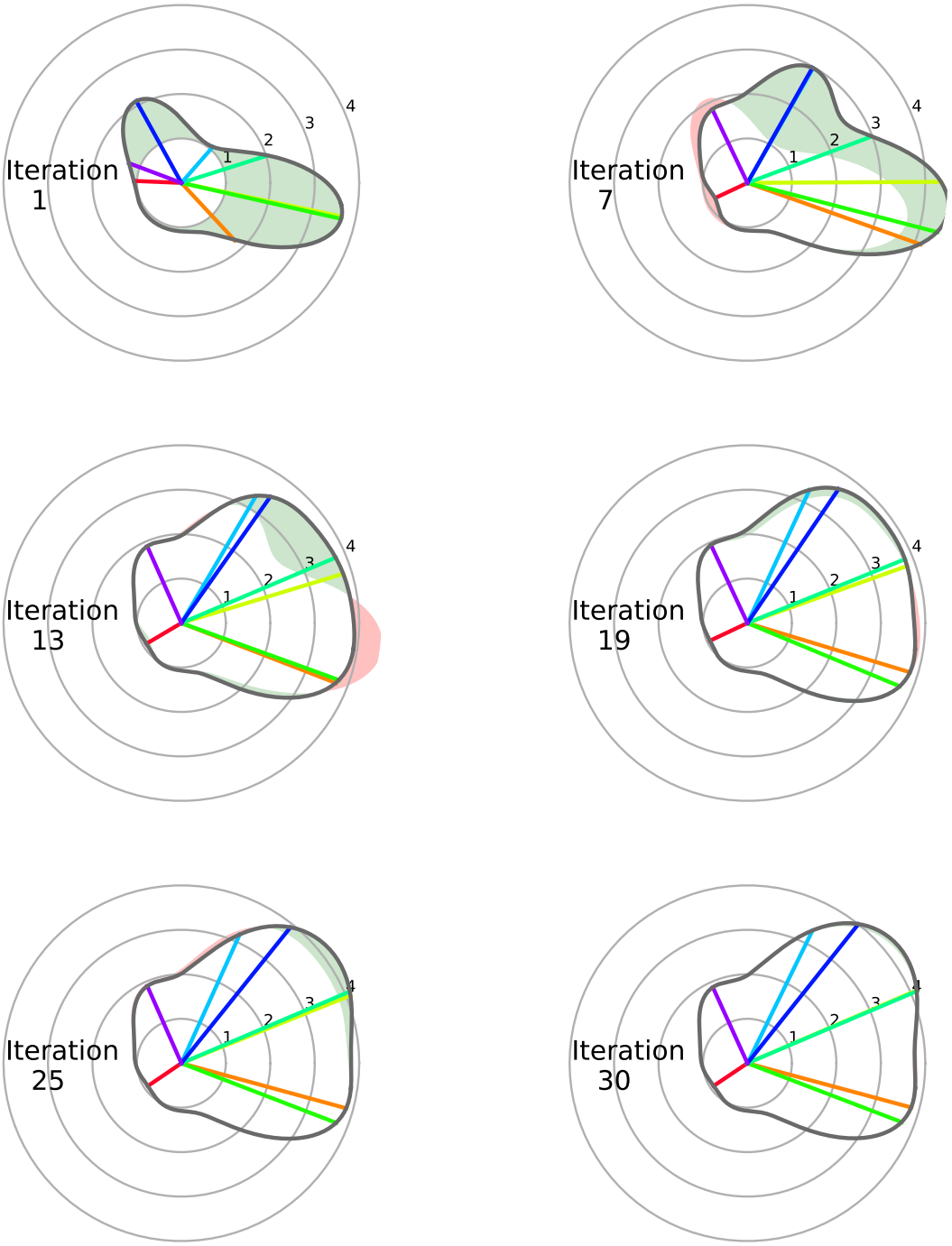

Figure 5.2: Visualization of the wind calibration function C_u (see 4.11) on select iterations of running RS+SGD.

Figure 5.3: Picture of the measurement station from above.

the movement of pollution at all.

The visual inspection showed similarities between the wind multipliers and the surrounding area. However, there is a need for a clear metric to justify that the model improves the wind measurements, and does not simply overfit to the data. Consequently, a trusted meteorological station was used to compare the results produced by the wind calibration, which can be seen in Figure 5.4. Finnish Meteorological Institute (FMI) has a weather station around 21 kilometers from the measurement station. Wind speed data from 2017 was gathered from their open data API [44]. Local weather data was averaged to 10 minutes, which represents the sample rate of the open data. For each iteration produced by RS+SGD, the correction multipliers were applied to local weather data for the whole year. Then RMSE was calculated against the data from FMI.

In Figure 5.4 RMSE improves quickly during the first few iterations and then starts to oscillate. The amount of oscillation gets lower likely because of the decreasing learning rate. It is apparent that at least in the first few iterations the model learns to correct wind speeds better. It should be noted that 0 RMSE is not necessarily perfect, since it would represent conditions 21 kilometers away in very different conditions, at 10 meters above the sea level.

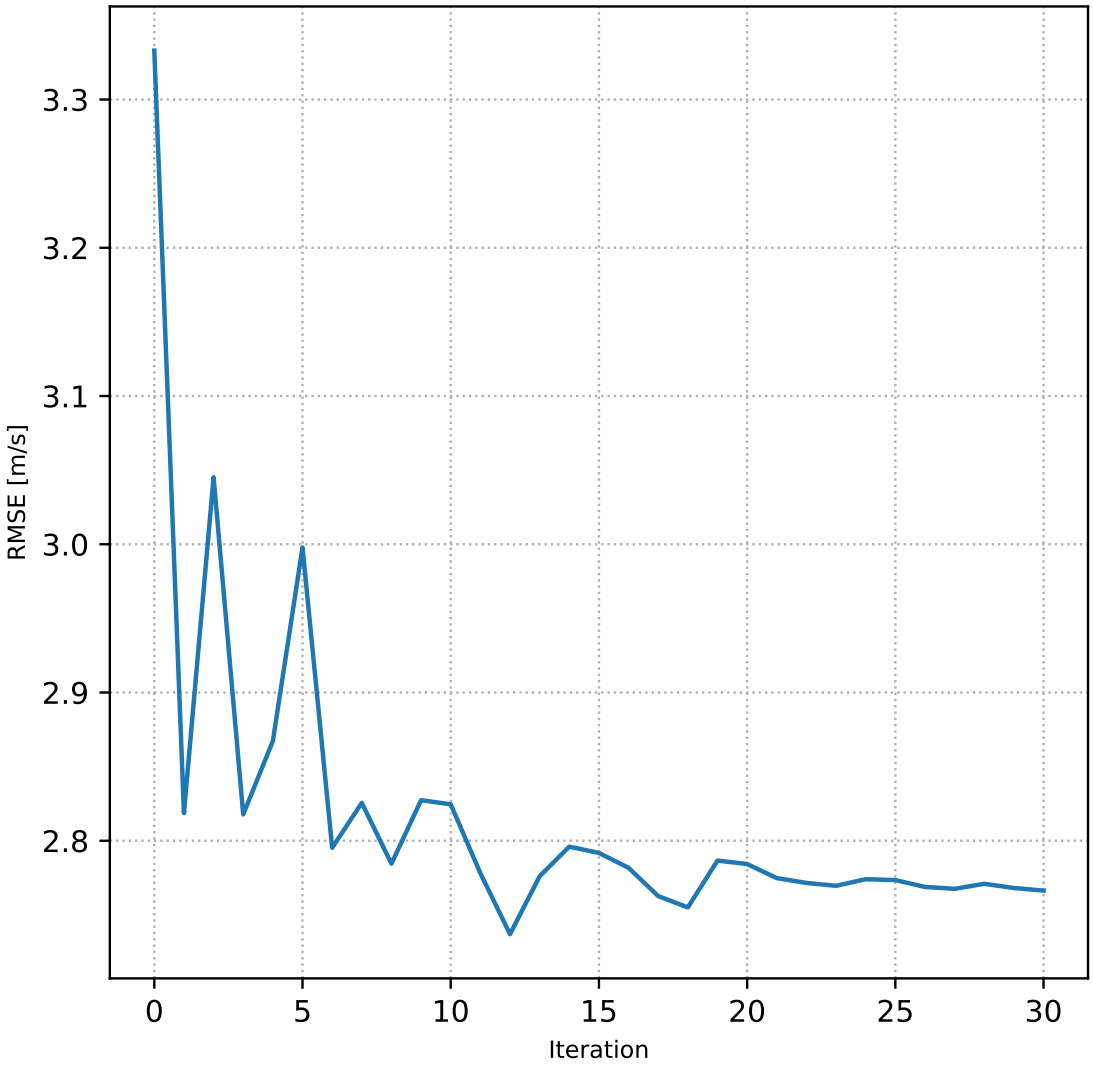

Figure 5.4: RMSE error per iteration of RS+GD when compared to data by FMI

Chapter 6

Conclusion

There are three main results in this thesis. Firstly, with very little control on the prediction and observation sets, a matching model can be optimized using simple classification metrics. This was done by adding the component of random performance estimation to used metrics. Secondly, matching can be done and improved upon without making estimates on ship emission rates using prior knowledge of ships. Thirdly, the model can optimize a wind calibration that improves the matching performance and which better represents the wind conditions between the station and the ship.

Random performance estimation was shown to make optimization possible when the prediction and observation sets can change almost arbitrarily based on model parameters. Moderating the effect of random matches seems to balance the results quite well with mostly a reasonable probability of random matches (< 0.05), even with the simple RS approach. But as was seen in table 5.2 for the case of training the model with only Class B ships, the optimal result may have quite a high probability of a random match as well. This suggests that simple subtraction of the expectation of random matches from true positives may not be enough in compliance monitoring, where a 10% chance of the result being essentially wrong is likely too much. This could of course be remedied by adding weighting to the expectation, or by adding a condition where parameters resulting in higher than a desired threshold probability of random matches are discarded.

While not directly observed in this thesis, the process of adding the expectation of random matches to performance metrics can prefer low probabilities too much. The expectation has the built-in assumption that the observed peaks are randomly distributed, which they are clearly not. It assumes the worst about the distribution of the data, and always reports an amount of random probability greater than zero. By design it gets harder to get lower random probability when the number of observations gets higher, which was seen in the comparison to the results in [20]. This may become a more pronounced problem in places with higher traffic density. This could be observed in the comparison to results in [20], where much looser parameters for Δt yielded a lower probability of random match due to the lower traffic density.

The results showed that in general matching both Class A and B ships using the same model decreased the performance compared to measuring Class A ships alone. The results indicate that they should not be matched using the same model, but maybe two different models or by using emission factor estimates.

Using iterative RS+SGD showed some increase in F_{1t}-score, but more importantly a -46.81 % relative decrease in the number of random matches. This result alone justified training the model for 30 iterations.

Iterative RS+SGD learned a wind calibration that visually matches the obstacles at the measurement station, and has a lower RMSE when compared to a weather station maintained by FMI. This information can both be used to improve the measurements at a station where no other weather stations are nearby, or to determine that instead a nearby weather station should be used.

The developed method is unbiased in the way that ships with stronger emission rates are treated the same as the ones with lower ones. The model optimizes for amount of correct matches independent of any prior information about ships (except of course when Class A and B ships were treated separately). It relies solely on the conditions of the measurement to decide if the ship's predicted peak is significant. This means that the

better the atmospheric conditions, the more likely it is that the prediction model estimates a predicted peak and therefore the more likely it is that it is also matched. This method is also independent of whether any information about the ship exists in a database at all. Unknown ships are likewise treated the same as known ships.

Large ships can be typically measured in worse conditions because the concentrations are much larger, but this does not necessarily mean that they should. Worse conditions may lead to longer travel times of the pollutant, and therefore more dry and wet deposition. Especially SO_2, which is used to determine FSC, reacts with water. The longer it has to interact with water, the lower the measured SO_2 concentration is. However, CO_2 does not react in this way, leading the ratio SO_2/CO_2 to be an underestimation of the FSC. This is why it is valuable to study the performance of unbiased models that do not estimate ship emission rates using prior knowledge of the ship. At least they should be studied to compare how much their performance can be improved by using a ship emission rate estimation model such as STEAM2. This thesis shows that prior information of ships is not strictly necessary to match them to measured pollution.

6.1 Open questions and future work

It is unknown what the performance of the model is for annotated data, where each ship and their pollution is annotated. To my knowledge, such studies have not been made. That method has the problem that it is not certain that humans would be better at annotating the data than models based on air pollution modeling. If annotator uses the same data as was available in this thesis, they would have to use some kind of air pollution model in the first place to estimate the correct peak from the data, which introduces many of the same model errors as are present in automatic matching. However, if visual confirmation of pollution travel would be available, the annotations could be very reliable. Annotated data naturally removes the need for random performance estimation altogether.

Peak detection was chosen based on its ease of implementation and runtime performance. Multiple alternatives exist that could offer better results. The current implementation of peak detection in AirNow, which is based on fitting curves to detect background concentration, was not compared against this method due to its complexity. However, the result of this thesis is a framework of testing the performance of multiple methods and their performance in matching easily without annotating data.

While the results of calibrating wind instruments using the data available were good enough to clearly see that they represent the installation conditions, it leaves open an area of additional research. The choice for the number of von Mises distributions and their initial concentrations was determined by only a few trials and left fixed. Other measurement stations should be tested using this method to see further validation results about the performance.

In this thesis performance metrics were only based on time of maximum concentration. The value of maximum concentration was only used for determining if the peak is to be discarded or accepted. The value of the maximum concentration could be used in a performance metric that values similar values for prediction and observation, for example by using RMSE. However, this would likely require a precise model for determining the emission rate of the ships (see 2.2.1). Additionally, it then requires a new method for determining the probability of a match happening at random.

Another question is the usage of the point of maximum concentration itself. This essentially loses a lot of information both the observations and predictions have. Prediction has multiple samples on a time series, where only the maximum value and its time is taken. Observations, with the help of a peak extraction (instead of detection) algorithm could have the same property. Comparison of the time series of peaks could provide additional improvement in the performance of the model without introducing ship emission estimates.

There are many areas in the pollution model that could receive further work. Most

of the choices were not carefully tested against other possible choices, due to the large variety of methods in the air pollution literature. They were merely tested to produce results that look good enough. Some of the hyperparameters, such as the decay rate or average stack height could be learned from data.

One of the major design choices was not to use a ship emission factor estimation model, such as STEAM2, to arrive at predicted peaks that also estimate the value of observed peaks. It would be interesting to see how this model would improve if such a model was included. Only then could it be decided if using such a model is warranted for better performance.

The generalization of the model was briefly discussed. Non-negligible amount of generalization error was present in all results between training and test sets. Iterative RS+SGD results showed that the error does not increase with further optimization, but is always present. Finding the reason for this is left as future work. This might require a separate validation and test set, where the validation set is used to minimize the generalization error. It should also be interesting to see how well the model generalizes to other measurement stations with the same hardware without the wind calibrations.

References

[1] Paolo Zannetti. *Air pollution modeling: theories, computational methods and available software*. Springer Science & Business Media, 2013.

[2] Development. Secretariat. *Review of maritime transport*. UN, 2018. VX.

[3] L. Kattner, B. Mathieu-Üffing, J. P. Burrows, A. Richter, S. Schmolke, A. Seyler, and F. Wittrock. Monitoring compliance with sulfur content regulations of shipping fuel by in situ measurements of ship emissions. *Atmospheric Chemistry and Physics*, 15(17):10087–10092, sep 2015. VT.

[4] International Maritime Organization (IMO). Faq. `http://www.imo.org/en/About/Pages/FAQs.aspx`, 2019. [Online; accessed 18-March-2019].

[5] Revised MARPOL Annex VI. Regulations for the prevention of air pollution from ships and nox technical code 2008. *International Maritime Organization*, 2009.

[6] Mikhail Sofiev, James J. Winebrake, Lasse Johansson, Edward W. Carr, Marje Prank, Joana Soares, Julius Vira, Rostislav Kouznetsov, Jukka-Pekka Jalkanen, and James J. Corbett. Cleaner fuels for ships provide public health benefits with climate tradeoffs. *Nature Communications*, 9(1), feb 2018.

[7] James J. Corbett, Paul S. Fischbeck, and Spyros N. Pandis. Global nitrogen and sulfur inventories for oceangoing ships. *Journal of Geophysical Research: Atmospheres*, 104(D3):3457–3470, feb 1999. V.

[8] James J. Corbett, James J. Winebrake, Erin H. Green, Prasad Kasibhatla, Veronika Eyring, and Axel Lauer. Mortality from ship emissions: A global assessment. *Environmental Science & Technology*, 41(24):8512–8518, dec 2007. V.

[9] Yewen Gu and Stein W. Wallace. Scrubber: A potentially overestimated compliance method for the emission control areas. *Transportation Research Part D: Transport and Environment*, 55:51–66, aug 2017. VTL.

[10] J-P Jalkanen, Anders Brink, Juha Kalli, Heidi Pettersson, Jussi Kukkonen, and Tapani Stipa. A modelling system for the exhaust emissions of marine traffic and its application in the baltic sea area. *Atmospheric Chemistry and Physics*, 9(23):9209–9223, 2009. VT.

[11] L. Johansson, J.-P. Jalkanen, J. Kalli, and J. Kukkonen. The evolution of shipping emissions and the costs of recent and forthcoming emission regulations in the northern european emission control area. *Atmospheric Chemistry and Physics Discussions*, 13(6):16113–16150, jun 2013. T.

[12] Pace Ralli. Industry insight: A survival guide for evaluating the cost of converting a vessel to use lng bunkers. `https://shipandbunker.com/news/features/industry-insight/566977-industry-insight-a-survival-guide-for-evaluating-the-cost-of-converting-a-vessel-to-use-lng-bunkers`, December 2015. [Online; accessed 18-March-2019].

[13] Ari Karppinen, Jari Härkönen, Juha Nikmo, Kari Riikonen, Jukka-Pekka Jalkanen, and Lasse Johansson. Compmon subactivity 1.2: Cost efficiency analysis.

[14] Tuomas Kokoi. Guidance for procuring sulphur monitoring services or equipment. techreport, Finnish Transport Safety Agency, October 2017.

[15] L Pirjola, A Pajunoja, J Walden, J-P Jalkanen, T Rönkkö, A Kousa, and T Koskentalo. Mobile measurements of ship emissions in two harbour areas in finland. *Atmospheric Measurement Techniques*, 7(1):149, 2014. TL.

[16] N. Berg, J. Mellqvist, J.-P. Jalkanen, and J. Balzani. Ship emissions of so DOAS measurements from airborne platforms. *Atmospheric Measurement Techniques*, 5(5):1085–1098, may 2012.

[17] J. M. Balzani Lööv, B. Alfoldy, L. F. L. Gast, J. Hjorth, F. Lagler, J. Mellqvist, J. Beecken, N. Berg, J. Duyzer, H. Westrate, D. P. J. Swart, A. J. C. Berkhout, J.-P. Jalkanen, A. J. Prata, G. R. van der Hoff, and A. Borowiak. Field test of available methods to measure remotely SO_x and NO_x emissions from ships. *Atmospheric Measurement Techniques*, 7(8):2597–2613, aug 2014.

[18] Petra Virjonen, Paavo Nevalainen, Tapio Pahikkala, and Jukka Heikkonen. Ship movement prediction using k-NN method. In *2018 Baltic Geodetic Congress (BGC Geomatics)*. IEEE, jun 2018.

[19] KINE Robot Solutions. Airnow emissions monitoring service. `https://airnow.fi/`. [Online; accessed 18-March-2019].

[20] Ari Karppinen, Jari Härkönen, Juha Nikmo, Kari Riikonen, Jukka-Pekka Jalkanen, and Lasse Johansson. Final report of compmon activity 4.4: Inversion tool.

[21] J-P Jalkanen, L Johansson, J Kukkonen, A Brink, J Kalli, and T Stipa. Extension of an assessment model of ship traffic exhaust emissions for particulate matter and carbon monoxide. *Atmospheric Chemistry and Physics*, 12(5):2641–2659, 2012.

[22] David B Stephenson. Use of the "odds ratio" for diagnosing forecast skill. *Weather and Forecasting*, 15(2):221–232, 2000. TL.

[23] John H Seinfeld. *Atmospheric chemistry and physics of air pollution*, volume 738. 1986.

[24] Albert Gyr and Franz-S Rys. *Diffusion and transport of pollutants in atmospheric mesoscale flow fields*, volume 1. Springer Science & Business Media, 2013.

[25] Gabriel T Csanady. *Turbulent diffusion in the environment*, volume 3. Springer Science & Business Media, 2012.

[26] Mark Z. Jacobson. *Fundamentals of Atmospheric Modeling*. Cambridge University Press, 2007.

[27] Wm J Veigele and James H Head. Derivation of the gaussian plume model. *Journal of the Air Pollution Control Association*, 28(11):1139–1140, 1978. T.

[28] D Bruce Turner. *Workbook of atmospheric dispersion estimates: an introduction to dispersion modeling*. CRC press, 1994.

[29] Spyros N. Pandis John H. Seinfeld. *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*. Wiley John + Sons, 2016.

[30] Jonathan D.W. Kahl and Hillary L. Chapman. Atmospheric stability characterization using the pasquill method: A critical evaluation. *Atmospheric Environment*, 187:196–209, aug 2018.

[31] Ferd Sauter, Margreet van Zanten, Eric van der Swaluw, Jan Aben, Frank de Leeuw, and Hans van Jaarsveld. The ops-model.

[32] Joseph S Scire, David G Strimaitis, Robert J Yamartino, et al. *A user's guide for the CALPUFF dispersion model*, January 2000.

[33] F.L. Ludwig, L.S. Gasiorek, and R.E. Ruff. Simplification of a gaussian puff model for real-time minicomputer use. *Atmospheric Environment (1967)*, 11(5):431–436, January 1977. TL.

[34] IMO SOLAS. International convention for the safety of life at sea. *International Maritime Organization*, 2003.

[35] ENTEC. Defra uk ship emissions inventory, final report. :http://uk-air.defra.gov.uk/reports/cat15/1012131459_21897_Final_Report_291110.pdf, November 2010.

[36] Thaddeus Vincenty. Geodetic inverse solution between antipodal points. *Richard Rapp Geodetic Science Ohio State University*, 1975.

[37] Girish Palshikar et al. Simple algorithms for peak detection in time-series. In *Proc. 1st Int. Conf. Advanced Data Analysis, Business Analytics and Intelligence*, volume 122, 2009.

[38] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, jun 2006.

[39] Kenneth J. Berry and Paul W. Mielke. A generalization of cohen's kappa agreement measure to interval measurement and multiple raters. *Educational and Psychological Measurement*, 48(4):921–933, dec 1988.

[40] Slavica Zec, Nicola Soriani, Rosanna Comoretto, and Ileana Baldi. Suppl-1, m5: High agreement and high prevalence: The paradox of cohen's kappa. *The open nursing journal*, 11:211, 2017.

[41] World Meteorological Organization. *WMO Guide To Meteorological Instruments And Methods Of Observation*, volume 8. 01 2008.

[42] José A. Carta, Celia Bueno, and Penélope Ramírez. Statistical modelling of directional wind speeds using mixtures of von mises distributions: Case study. *Energy Conversion and Management*, 49(5):897–907, may 2008.

[43] Leon Bottou. *Stochastic Gradient Descent Tricks*, volume 7700 of *Lecture Notes in Computer Science (LNCS)*, pages 430–445. Springer, neural networks, tricks of the trade, reloaded edition, January 2012.

[44] Finnish Meteorological Institute (FMI). Open meteorological data. https://en.ilmatieteenlaitos.fi/open-data.

Appendix A

Additional results

	Training		Test		
	a	ACC_t	a	ACC_t	$Pr(X^t)$
Class A	666	0.369	556	0.331	0.052
Class B	115	0.157	110	0.142	0.129
Both	724	0.337	633	0.311	0.038
Class A (year)	2204	0.379	2054	0.364	0.044
Both (year)	2136	0.364	2002	0.346	0.04

Table A.1: The results of running RS when filtering routes by their AIS Class, maximizing true accuracy.

	Peak observed	No observation	Total
Peak predicted	717	275	992
No prediction	805	0	805
Total	1522	805	1267

Table A.2: Contingency table for Class A ships in summertime result

	Peak observed	No observation	Total
Peak predicted	97	150	247
No prediction	258	0	258
Total	355	258	397

Table A.3: Contingency table for Class B ships in summertime

	Peak observed	No observation	Total
Peak predicted	752	424	1176
No prediction	885	0	885
Total	1637	885	1600

Table A.4: Contingency table for both AIS classes in summertime

	Peak observed	No observation	Total
Peak predicted	2204	684	2888
No prediction	2548	0	2548
Total	4752	2548	3572

Table A.5: Contingency table for Class A ships in the whole year

	Peak observed	No observation	Total
Peak predicted	2197	962	3159
No prediction	2456	0	2456
Total	4653	2456	4121

Table A.6: Contingency table for both AIS classes in the whole year

	Peak observed	No observation	Total
Peak predicted	2183	770	2953
No prediction	1926	0	1926
Total	4109	1926	3723

Table A.7: Contingency table using alternating RS and SGD for Class A ships in the whole year