Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Matevž Ropret

# Napovedovanje koncentracij onesnaževalcev zraka in prepoznavanje izvornih regij

magistrsko delo

študijski program druge stopnje
računalništvo in informatika

Mentor: doc. dr. Erik Štrumbelj

Ljubljana, 2016

University of Ljubljana
Faculty of Computer and Information Science

Matevž Ropret

# Forecasting air pollutant concentrations and identifying source regions

master's thesis

second level program of
computer and information science

Mentor: assist. prof. Erik Štrumbelj

Ljubljana, 2016

# Povzetek

**Naslov:** Napovedovanje koncentracij onesnaževalcev zraka in prepoznavanje izvornih regij

Onesnaževalci zraka lahko predstavljajo velik problem za zdravje ljudi. Onesnaževalci se lahko prenašajo z gibanjem zračnih mas iz izvorne v druge regije. Zanima nas, ali lahko s pomočjo gibanja zračnih mas napovemo, kakšna bo koncentracija onesnaženosti za nek dan in ali lahko ugotovimo, od kod prihaja onesnaženost. Obstaja že veliko literature na to temo, razvitih pa je bilo tudi nekaj metod za reševanje tega problema. Mi smo želeli uporabiti strojno učenje, da bi naredili nove, boljše metode. Naredili smo dve novi metodi. Prva temelji na principu 2D mreže, druga pa neposredno uporabi kar koordinate o trajektorijah gibanja zračnih mas. Ti dve metodi smo primerjali z obstoječima metodama CF in RCF. Končni rezultati so pokazali, da so nekatere naše metode vsaj dvakrat boljše pri napovedovanju onesnaženosti. Vseeno končni rezultati niso tako dobri, kot smo pričakovali. Na vizualizacije, od kod prihaja onesnaženost, se ne moramo preveč zanesti, saj nam je uspelo vizualizirati samo metodo 2D mreže, ki pa ne daje boljših rezultatov od obstoječih. Kombinirali smo tudi rezultate večih postaj v upanju, da bi to izboljšalo vizualizacijo, ampak tudi ta pristop ni dal boljših rezultatov.

**Ključne besede:** strojno učenje, vir onesnaženosti, CF, RCF, napovedovanje onesnaženosti, naključni gozdovi, blasso, bayesovska regresija.

# Abstract

**Title:** Forecasting air pollutant concentrations and identifying source regions

Air pollutants are hazardous to human health. Pollutants can be transported by air masses from one region to another. We were interested if we could use air mass movement to predict daily pollution concentrations and to visualize where this pollution came from. This area is rich in related work and there already exist methods that solve this problem. Our goal was to use machine learning to create new and better performing methods. We created two new methods. The first is based on a 2D grid, while the second is based on raw coordinate data. We compared these two methods with existing CF and RCF methods. Results show that some of our methods perform more than twice as good as existing methods. However, the results are still below our initial expectations. We cannot rely on source attribution visualization, because we were able to get it working only with the 2D grid method, which is not much better than existing CF and RCF methods. We also tried combing results of multiple stations in hopes that we could make better source attribution visualization, but this also performed worse than expected.

**Keywords:** machine learning, pollution source, CF, RCF, source attribution, pollutant forecasting, pollutant prediction, random forest, blasso, bayesian regression.

*Zahvaljujem se mami Zdenki, ki mi je pomagala skozi čas študija. Zahvaljujejm se tudi mentorju Eriku Štrumblju, ki se je izjemno posvetil mentorstvu in Greti Gašparac, ki je prispevala k pridobivanju podatkov EMEP in implementaciji algoritma RCF.*

# Contents

# List of notation and shorthands

| **Shorthand** | **Meaning** |
|---|---|
| **TSM** | Trajectory Statistical Methods |
| **CF** | Concentration Field method |
| **RCF** | Redistributed Concentration Field method |
| **RF** | Random Forest algorithm |
| **blasso** | Bayesian LASSO regression |
| **XYZ** | the raw trajectory transformation method |
| $l$ | trajectory index |
| $c_l$ | pollution of trajectory $l$ |
| $d$ | day index |
| $n_d$ | number of days |
| $c_d$ | pollution of day $d$ |
| $\beta_{ij}$ | pollution intensity (concentration) of cell $ij$ |
| $\beta_{ij}^0$ | initial concentration field value of cell $ij$ |
| $n$ | number of trajectories |
| $n_l$ | number of segments for trajectory $l$ |
| $B_{lk}$ | pollution of the grid cell in which the $k$-th segment of trajectory $l$ is in |

# Razširjen povzetek v slovenskem jeziku

Onesnaženost zraka z majhnimi delci ali strupenimi plini povzroča razne bolezni, kot so omotica, znižana odpornost na okužbe, bolezni dihal in srčno-žilne bolezni. Zaradi tega se je veliko strokovnjakov posvetilo raziskovanju onesnaženosti zraka. Viri, ki prispevajo k onesnaženosti neke regije, so lahko lokalni ali pa zelo oddaljeni, saj se onesnaženost lahko prenaša preko gibanja zračnih mas. V tem delu se posvečamo odkrivanju izvornih regij onesnaženost in napovedovanju koncentracije onesnaževalcev v zraku. Natančneje, delo se osredotoča na napovedovanje onesnaževalcev, za katere imamo na voljo pogoste meritve (dnevne) in gibanja zračnih mas v obliki 2-dimenzionalnih ali 3-dimenzionalnih trajektorij, ki potekajo od receptorja. Obstaja veliko sorodnih del na to temo (glejte Fleming et al. [6]). Skoraj vsa sorodna dela spadajo v eno od dveh skupin. V prvi skupini so dela, kjer položijo 2-dimenzionalno mrežo celic čez določeno ozemlje. Onesnaženost v zraku se potem pripiše celicam, skozi katere so potovale trajektorije (onesnaženost ali enakomerno razporedimo po celicah ali glede na čas, ki ga trajektorije preživi v celici). Najbolj pogosti metodi sta PSCF (Potential Source Contribution Function) iz Ashbaugh et al. [2] in CF (Concentration Field) metoda iz Seibert et al. [29]. Glavni problem teh dveh metod je napačno prikazovanje onesnaženosti, saj se onesnaženost doda vsem celicam, skozi katere potuje trajektorija, in ne samo dejanskimi izvornim regijam. Ta problem delno reši metoda RCF (Redistributed Concentration Field) iz Stohl [34]. Druga skupina metod

uporablja algoritem gručenja, kjer trajektorije najprej uvrstimo v gruče glede na njihovo pot. Te gruče se potem uporabijo kot atributi za modele oz. bolj pogosto le primerjamo koncentracije onesnaženosti med gručami. Ta pristop lahko grobo oceni izvor onesnaženosti, ne moremo pa določiti natančnih regij, zato teh metod nismo vključili v delo. Bolj natančen opis sorodnih del je v poglavju 1.1.

Naš problem si lahko interpretiramo kot problem napovedovanja - želimo napovedati koncentracijo onesnaženosti v zraku s pomočjo trajektorij zračnih mas. Dosedanje metode se zelo preproste in smo mnenja, da lahko z metodami strojnega učenja in statistike z minimalnim trudom naredimo boljše (ne-linearne) napovedne modele (glejte poglavje 2). Naredili smo tudi preprosto metodo, ki uporabi podatke večih postaj hkrati. Kompleksnost takih modelov je problematična, saj večino modelov strojnega učenja ni enostavno interpretirati, kar pomeni, da je določanje izvornih regij lahko oteženo ali celo nemogoče. To rešimo z metodo, predlagano v [36, 37]. Ta metoda model obravnava kot črno škatlo, kar pomeni, da lahko uporabimo katerikoli model strojnega učenja za učenje in razlago izvornih regij. Ključno vprašanje za te nove metode strojnega učenja je, kako naj transformiramo vhodne podatke, da bodo v obliki, iz katere se lahko model nauči največ. Kot smo opisali v poglavju 2, smo uporabili dva različna pristopa: prvi pristop uporablja mrežo celic, drugi pa kar neposredne uporabi koordinate trajektorij. V poglavju 3 opišemo empirično evaluacijo naših novih in obstoječih modelov na približno 15 letih podatkov (se razlikuje od postaje do postaje) iz realnega sveta za več različnih vrst onesnaževalcev ($PM_{10}$, $SO_2$, $NO_2$, $O_3$, $PM_{2.5}$), ki jih meri več postaj po Evropi. Kljub veliki količini sorodnih del obstaja zelo malo sistematičnih kvantitativnih primerjav. Scheifinger and Kaiser [27] je primerjal PSCF, CF, and RCF v idealiziranih pogojih (virtualne trajektorije in viri onesnaženosti), in v pogojih podobnem realnem svetu (primerjal je z inventarjem emisij onesnaževanja). Ugotovil je, da metode delujejo dobro v idealiziranih pogojih (predvsem RCF), ampak pod realnimi pogoji pa delujejo slabo. Brereton et al. [3] so ugotovili enako. Kong et al. [15] in Ying-

Kuang Hsua [39] so z uporabo inventarja emisij onesnaževanja in z vizualno inšpekcijo ugotovili, da RCF deluje boljše kot PSCF, saj nima repa.

Rezultati empirične evaluacije, predstavljeni v poglavju 4, so nas razočarali, saj smo pričakovali, da bodo naši novi modeli delovali veliko bolje. V poglavju 5 smo povzeli naše rezultate in napisali potencialne izboljšave za prihodnje raziskave. Naredili smo 3 različne teste. Prvi test predpostavi, da vsa onesnaženost prihaja iz ene točke. Ta test vse metode opravijo dobro. Drugi test uporabi 4 točke kot vire onesnaženosti. Naše metode se tukaj izkažejo za boljše. Tretji test uporabi bolj realen vir onesnaženost, saj uporabimo bolj razmazano regijo kot vir onesnaženosti. Tukaj metode ne delujejo preveč dobro, kar predstavlja velik problem, saj je ravno ta test najbolj kritičen. Uporaba podatkov z več postaj hkrati prav tako ne izboljša natančnosti vizualizacije izvornih regij.

Uporabili smo podatke iz petih postaj (Iskrba, Zingst, Illmitz, Svratouch in Westerland). Pri treh postajah (Illmitz, Westerland in Zingst) napovedujemo onesnaženost dvakrat bolje kot pri ostalih dveh postajah, česar nam ni uspelo pojasniti. Želeli smo tudi ugotoviti, kateri onesnaževalec lahko najbolje napovemo. Vemo, da se delci ($PM_{10}$, $PM_{2.5}$) veliko lažje raznašajo po zraku z gibanjem zračnih mas, kot pa plini ($SO_2$, $NO_2$, $O_3$). To so potrdili tudi naši rezultati v tabeli 4.1. Rangirali smo tudi vse metode in modele po uspešnosti napovedovanje, kar lahko vidite v tabeli 4.7. Nekateri naši modeli so bili dvakrat boljši od obstoječih, a ravno našega najboljše delujočega modela nam ni uspelo vizualizirati. Vizualizacijo izvornih regij na resničnih podatkih lahko vidite na slikah 6.11 in 4.4. Rezultate uporabe podatkov z več postaj si lahko ogledate na slikah 6.9 in 6.10. Dobljeni rezultati se ujemajo z znanimi podatki: vemo, da na $O_3$ zračne mase skoraj ne vplivajo, saj tudi vsi modeli ta onesnaževalec napovedujejo veliko slabše od ostalih. Delce $PM_{10}$ in $PM_{2.5}$ pa se napoveduje veliko boljše, kar se tudi ujema s sorodnimi deli. Model RF (random forest) je vedno dal boljše rezultate kot model blasso (Bayesovska različica L1 regresije). Obe naši novi metodi sta boljši od obstoječih, XYZ še posebno, a te metode nam ni uspelo vizualizirati. Zelo

problematična je slaba napovedna moč CF in RCF metode, saj ju uporablja ogromno obstoječih del, kar vzbudi dvom v njihove zaključke.

Potrebno je omeniti, da so trajektorije lahko zelo nenatančne v nekaterih pomembnih regijah, a žal ne moremo vedeti, v katerih. Uporabljene trajektorije niso bile izračunane na meteorološkem polju najvišji ločljivosti. Mogoče bi bili rezultati boljši, če bi uporabili bolj natančne trajektorije in hkrati uporabili polje celic, ki prekriva manjše območje, saj onesnaževalci bližje merilni postaji bolj vplivajo na dnevno onesnaženost, kot bolj oddaljeni onesnaževalci. Prav tako se verjetnost napake v trajektoriji povečuje z njeno oddaljenostjo od merilne postaje.

Na sliki 4.4 vidimo nekaj, kar bi lahko bilo zelo pomembno: naša metoda mreže celic z RF in blasso modelom prikaže povsem drugačno vizulizacijo kot metodi CF in RCF. Naša metoda označi vire onesnaženosti bližje postaji, medtem ko CF in RCF trdita, da so viri bolj oddaljeni. To lahko pomeni, da naša metoda zazna lokalne regije blizu postaje, ki so bolj pomembne pri prenosu onesnaževalcev (npr. zaradi vpliva reliefa, gozdov...), medtem ko pa CF in RCF zaznata bolj grobe regije, od kjer se onesnaženost dejansko prinese. To bi pojasnilo, zakaj naše metode dajejo boljšo napoved za dnevno koncentracijo onesnaženosti, saj bolj upoštevajo celice blizu postaje, od koder je večja verjetnost, da onesnaževalci dejansko pridejo. Rezultati postaje Iskrba se skladajo s sorodnimi deli, saj vemo, da v Slovenijo največ onesnaženosti pride iz severo-vzhoda in juga.

Ker vse metode dajejo relativno slabe rezultate, je nemogoče reči, ali slike prikazuje dejansko pomembne regije ali pa gre za dejanski šum. Metoda uporabe podatkov več postaj ne izboljša rezultatov. Če bi naše metode (vključno z CF in RCF) uporabili meteorologi, bi priporočali, naj uporabijo kombinacijo CF, RCF in metode mreže celic z RF modelom in naj rezultate CF in RCF uporabijo za analizo grobih, oddaljenih virov onesnaženosti, medtem ko bi metoda mreže celic z RF modelom bila uporabna za analizo regij v bližini postaje. Predvidevamo, da so za slabo napovedno moč lahko krivi naslednji razlogi:

- Nenatančne trajektorije.
- Sezonske spremembe gibanja zračnih mas, kar pomeni drugačno pot trajektorij skozi leta.
- Spremembe v naravi in okolici (gradnja tovarn, urbanizacija, sprememba gozdne površine...).
- Onesnaževalci se manj prenašajo preko zračnih mas, kot smo mislili. Če je to res, potem so vsi modeli dokaj neuporabni.
- Ne upoštevamo disperzije in turbulenc v zraku, kar lahko povzroči zelo drugačno gibanje delcev. To predvidevata Scheifinger in Kaiser [27].

Za prihodnje raziskave priporočamo uporabo bolj natančnih trajektorij in hkrati uporabo mreže celic, ki pokriva manjše, bolj lokalno območje. Učenje na podatkih iz več let je lahko problematično zaradi sezonskih vplivov, zato bi modeli morali to upoštevati. Prav tako bi modeli morali upoštevati disperzijo in/ali turbulentno gibanje delcev v zraku. Vizualizacija metode, ki koordinate uporablja neposredno, bi lahko bolj natančno odkrila izvore onesnaženosti, saj napoveduje vsaj dvakrat bolje kot obstoječe metode. Bolj kompleksen model, ki uporablja podatke z več postaj hkrati, je tudi pomembno izhodišče za prihodnje delo.

Nekateri rezultati tega dela so bili objavljeni v [25].

# Chapter 1

# Introduction

Air pollution in the form of fine particles or noxious gasses has been linked with a wide range adverse health effects, such as fatigue, reduced resistance to infection, respiratory diseases, and cardiovascular diseases. As a result, a lot of expert and research effort is being dedicated to the study and the management of air pollution. Sources that contribute to air pollution concentration levels at a particular location (receptor) can be local or potentially very distant, due to the process long-range transport of atmospheric pollutants in an air mass. In this thesis we focus on the problem of identifying potential source regions from concentration measurements and air mass transport information. In particular, we focus on the most common variant of this problem, where source regions are to be identified given periodic (typically daily or hourly) concentration measurements over a longer period of time and corresponding air mass transport information in the form of 2-dimensional or 3-dimensional back-trajectories from the receptor location. Related work on this problem is very rich with applications (see Fleming et al. [6] for a review). In terms of methodology, almost all related work belongs to one of two general groups of methods. The first group are methods based on a grid tessellation of the area of interest. Concentration levels or high-concentration episodes are then attributed to (typically equally or according to residence time) all grid cells passed by the corresponding trajectory. The most common

variants are the PSCF (Potential Source Contribution Function) method of Ashbaugh et al. [2] and the CF (Concentration Field) method of Seibert et al. [29]. The main issue with these methods is that a grid cell may be (falsely) identified as a potential source only because it is often passed by trajectories that also pass through actually polluted regions (that is, a trailing effect). This was partially addressed by the reweighting approach RCF (Redistributed Concentration Field) by Stohl [34]. The second group are clustering-based approaches, where first a clustering algorithm is used on the trajectories to divide them into distinct clusters (according to their paths) and then the cluster membership is used as a discrete variable for predicting concentration levels or, more typically, compare clusters on concentration levels. These approaches are suitable for identifying general polluted airmass pathways, but not for identifying specific locations of potential sources and will therefore not be considered in the comparison. We provide a more detailed description of the key related work in Chapter 1.1.

The problem of interest is in essence a prediction problem - we wish to predict concentration levels from back-trajectory data. Given the relative simplicity of the methods used in related-work, we hypothesized that we could potentially achieve substantially better results with minimal effort by drawing on the vast machine-learning and statistical prediction toolbox and using more complex (non-linear) prediction models (see Chapter 2 for details). However, although this might lead to more accurate predictions (and in turn a more accurate model for potential source regions), the resulting models are typically complex, which makes them difficult to interpret or to extract a meaningful identification of potential source regions. We deal with this by using a black-box approach to computing input variable contributions for a prediction model proposed in [36, 37]. The key issue with such an approach is how to transform the raw back-trajectory path data into a form that is suitable for learning/fitting models. As described in Chapter 2, we use two different approaches: grid-based tessellation and raw trajectory path coordinates. In Chapter 3 we describe our empirical evaluation of the pre-

dictive quality of classic and proposed approaches on approximately 15 years (varies from station to station) of real-world data for multiple particulate matter concentrations ($PM_{10}$, $SO_2$, $NO_2$, $O_3$, $PM_{2.5}$) as measured at multiple monitoring stations across Europe. Despite the widespread application of source attribution methods, there have been few systematic quantitative comparisons. Scheifinger and Kaiser [27] compared the PSCF, CF, and RCF in an idealized setting (virtual source volumes and real trajectories), and a real-world setting (comparison to pollution emission inventory) and found that the methods (RCF in particular) work well in an idealized setting, but not in a real-world setting. Brereton and Johnson [3] also used simulations and also found that RCF was the best at identifying source regions. Kong et al. [15] using emission inventory data and Ying-Kuang Hsua [39] using visual inspection of results also found that RCF is slightly better and does not feature the trailing effect common in PSCF.

The results of the empirical comparison, which are presented separately in Chapter 4 are disappointing, as we assumed that our new models would work much better. In Chapter 5 we summarize our findings and contributions and provide some directions for future work.

## 1.1 Related work

While this topic is very rich in related work many of them describe the same approach by using a different name, thus making it a confusing area to research. In this following Subsections we will present the most important related work and methods.

### 1.1.1 Overview

Scheifinger and Kaiser in [27] compare different methods in a controlled virtual environment: PSCF (Potential Source Contribution Function, cannot be used for prediction thus making it useless for our case), CF, RCF. They call these methods trajectory statistical methods (TSM). They argue that we

don't know well enough how good/bad these methods perform in complex situations. They use virtual sources to test for specific cases alongside with real data. PSCF and CF underestimate high emission areas and overestimate low emissions areas. RCF is much better compared to these two. They conclude that in idealized cases all methods perform reasonably and are usable. Explained spatial variance is used to measure performance. The performance on idealized data is 60-80%, while on real data it's only 15% for the total area or 20-30% for at a spatial coverage of about 60-70%. Here PSCF is better than the other two methods. By using a more complex formula, they compute the difference between the real data and predicted data. This shows where there are potential over- or underestimations of sources. Turbulent dispersion and removal processes are neglected by TSMs. These two are the most likely sources of the decreased performance in real data compared to simulated. They introduce a decay function which simulated turbulence and dispersion and apply it to virtual data. This produces similar results to real data so they conclude that turbulence and dispersion are indeed factors, but further studies would have to be done to show how this affects TSMs. In [7] they compare 3 different methods: statistical metrics/comparison, concentration field and cluster analysis. Hsu et al. [39] discover that a combination of multiple approaches give better results. In [15] they use a two-stage cluster and compare it to self-organizing maps(SOM). Then they use PSCF and RCF. SOM with Mahalanobis metric proved to be better for clustering. Combined PSCF and RCF gave better results. To combine them they used the average of normalized values of PSCF and RCF. Kaiser et al. [12] explain that CF and RCF depends on concentration, therefore seasonal variation may cause problems, They don't do any real comparison. In [1] we can see an overview of statistical and back-trajectory dispersion methods. In [3] we find a very good overview of PSCF, CF, QTBA (quantative transport bias analysis), RCF and also compares them. Using RCF alongside CF seemed best. RCF, CF and PSCF work well in areas with large trajectory coverage. QTBA did not perform well. RCF was best, but gave false source regions. RCF was

improved when combined with CF.

### 1.1.2 Other related work

Kaiser et al. in [12] state that CF and RCF depend on concentration, therefore seasonal variation may cause problems. They don't do any real comparison. An overview of statistical and back-trajectory dispersion methods is in [1]. A very good overview of PSCF, CF, QTBA, RCF and their comparison is in [3]. Using RCF alongside CF seemed best. RCF, CF and PSCF work well in areas with large trajectory coverage. QTBA did not perform well. RCF was best, but gave false source regions. RCF was improved when combined with CF. A virtual simulation is performed in [5]. They recommend using data from multple stations. A review of source appointment research for particulate matter ($PM_{10}$ and $PM_{2.5}$) is in [38]. They also state that the combination of back-trajectory and and source apportionment analysis has much potential. These papers use clustering and then a statistical analysis, but these methods are not suitable for detecting source regions: [13, 15, 18, 19]. 3D clusters are used (minimum convex hull of clustered back-trajectories) to separate low and high air mass flows in [19] uses. A review of related work can also be found in [6]. A better method for computation of backwards trajectories is presented in [35]. The backwards trajectory is not a single line but turns into a filamentary structure because of turbulence and convection. The Lagrangian particle dispersion model is used with cluster analysis of particle positions to derive better "trajectories" and trajectory ensembles. This reduces error by filamentation and backwards growth. In [17] they add an exponential term to the residence time analysis for the probability that the pollutant won't be carried all the way from the source to the receptor. Another paper that uses PSCF and CF is [8], where they confirmed increased pollution from heavily polluted areas to the measuring station. Pinxteren et al. [22] use chemical source apportionment, not spatial. They concluded that PSCF cannot distinguish large source from moderate ones because of the criterion at which we specify if a trajectory is polluted or not, see [15].

CF and RCF was developed to fix this. The method proposed in [2] is arguably the most commonly used source attribution method. The method is based on a cell-based tessellation of the area of interest. Raised concentration levels are attributed to participating cells, according to trajectory residence time, typically estimated as the number of backward trajectory vertices that fall within the cell. Many subsequent papers use this method, sometimes with modifications: [4] and [40] use a simple weighting, where cells with a smaller number of trajectories are weighted less, to ensure statistical stability. Salvador et al. [26] apply a binomial statistical test to each cell. In [28] they decompose RT into two fields: transport directional frequency and the inverse transport speed. The spatial patter in default RT method and transport direction frequency are very similar: from this they conclude that RT works predominantly based on the frequency at which the trajectories traverse the given area before they reach the receptor.A summary of [29] is in [34]. The problem with [29] technique is that it assumes equal distribution of the pollution along the trajectory. In reality, pollution sources are concentrated on a small area along the trajectory. The method by [34] (RCF) fixes this problem. An improved method presented by [29] is in [34], where it is called RCF. In [24] we see a variant of the concentration field method: they add a weight factor that represents the trajectory inaccuracy. This inaccuracy factor was obtained from other literature that computed trajectories and their inaccuracies. See [6] for identification of regions which are more or less likely to be traversed during high or low concentrations during the day. This work also contains a combination of trajectory studies with source apportionment models and clustering. Hsu et al. [39] use PSCF, CF and residence time weighted concentration. A modified version of these methods is also discussed. These methods are combined to give better predictions. PSCF and CF appear to be able to distinguish between moderate and large sources. RCF resolved high potential source area. RCF combined with JP-PSCF (joint probability PSCF) removed the tailing effect that happens with pure PSCF. Flexpart is a model used to compute particle dispersion based on

physical methods. By running it backwards we can trace particles from receptor to source regions. See [32] for description of FLEXPART. See [31] for an example of this. Residence time is used. They say that FLEXPART model makes trajectories obsolete. Compared to the trajectory model, FLEXPART model takes in account the growth of the retroplume (the potential source field computed by backward simulation). The inversion algorithm adjusts the emissions used in the model to better match the observed and simulated concentrations (see [33], they use inversion method based on [30]). In [16] we see a proposal of using multiple receptors simultaneously to better identify sources. Artificial data were generated. They compared single and multi receptor models and showed that multi receptor model is better. They use condition probability (CP) which works better for direction rather than actual location prediction. QTBA is introduced in [14]. Multiple probability functions are multiplied and integrated over to produce a probability (density) field. None of the probability functions are well defined and are purely approximations of real world processes (dispersion, reaction, deposition, ...), therefore it is a hybrid approach as it uses back-trajectories alongside models of chemical processes. In [41] QTBA is compared with RCF. They develop Simple QTBA (SQTBA) which just ignores the effects of chemical reactions and depositions in the probability function. The results show that SQTBA and RCF clearly identify large and clearly defined sources. The tailing results of SQTBA (the proposed how to reduce it) can identify false source areas. SQTBA gives reliable results at lower spatial resolution. RCF gives better spatial resolution and can detect small hot spots but it misses a lot of source areas. RCF is sensitive to the influence of by many factors (deposition, reactions, variation in emission rate). RCF must be used with caution when there is high variation in emission rate. QTBA is also used in [9], where they use it to identify pollution cased by coal usage. In [10] they combine QTBA with PMF (positive matrix factorization, used to identify source-receptor relationship based on chemical composition), see [9]. In [11] we see that using box tessellation introduces sampling distortions. We can use proba-

bility density function centered around each data point to estimate spatial statistics. They show some existing methods to compute kernel estimation on a sphere and it then proposes a new, faster method. The final results of this method compared to tessellated box methods should not differ greatly. The main difference between [20] and most other papers is that they don't use a grid cell but density functions (kernels), it is based on [11]. The goal of this thesis is the "origin averaging technique": to estimate relationship between origin of trajectories and the concentration measured. In [27] they compare trajectory statistical methods (TSM): PSCF, CF and RCF. The results say that the transport processes is very simplified by the trajectory model and this causes inaccuracies. They say that TSM methods should not be trusted in general. They conclude that high emission area are underestimated, low emission areas are overestimated. The real world data simulation (cloudy/noisy sources) performed a lot worse than the point-source idealized version. Adding an exponential term made test data on idealised point sources gives results similar to real world data simulation. From this we can conclude that dispersion and deposition are indeed the source of the error between the idealised point data and real world cloudy/noisy data.

## 1.2 A detailed look at CF and RCF

The data in our problem setting are composed out of two parts. The first part is a sequence of pollution concentration measurements $c_d > 0, d = 1..n_d$ for each day $d$ for a particular pollutant at a particular monitoring station.

The second part are trajectories of air mass movement to the station's location for the period during which the concentrations were measured. Each air mass trajectory is a path consisting of sequential vertices, where each vertex can be seen as a 3D coordinate (latitude, longitude, altitude), although only latitude and longitude coordinates are used in the methods described in this thesis (altitude is ignored). All trajectories in a day correspond to one daily pollution measurement.

Note that all trajectories for day $d$ are assigned the same pollution value, that is if day $d$ pollution measurement is $c_d$, we set the pollution $c_l = c_d$ to each trajectory in the trajectory set $l$ which belongs to day $d$.

Concentration field (CF) and redistributed concentration field (RCF) are based on a grid tessellation of the area of interest. Concentration levels are then attributed to grid cells passed by the corresponding trajectory. The most common is the CF method. The main issue with CF is that a grid cell may be misidentified as a potential source because it is often passed by trajectories that also pass through actually polluted regions. This was addressed by the re-weighting approach RCF.

Note that CF is also known as concentration weighted trajectory (CWT). RCF is also known as residence time weighted concentration field (RTWC).

The transformation applied to trajectories before the application of CF or RCF methods is grid-based. The geographical area of interest is tessellated with a rectangular grid and each trajectory $l$ is transformed into a matrix of residence times $t_{lij} \geq 0$, one for each cell $ij$. Residence time tells us how much time a trajectory spent in that cell. Because each trajectory is described with a set of vertices, we estimate residence times in individual cells by assuming a straight path with constant speed between adjacent (in time) vertices. We compute this by stepping along subdivided trajectory segments.

### 1.2.1 Concentration Field

The CF method starts with a grid-based transformation of trajectories. The intensity $\beta_{ij}$ of the pollution sources in grid cell $ij$ is then calculated as

$$\beta_{ij} = \frac{\sum_{l=1}^{n} t_{lij} c_l}{\sum_{l=1}^{n} t_{lij}}, \tag{1.1}$$

where $c_l$ is the pollutant concentration associated with the $l-$th trajectory and $t_{lij}$ is the $l-$th trajectory's residence time in cell $ij$.

The CF method is used exclusively for visualization of cell intensities and visual inspection of potential sources of pollution. However, it is in essence

a model that connects trajectories and concentrations and could be used to infer the latter from the former. In the following, we extend the approach to predicting concentrations.

In order to derive a prediction for a (possibly new) trajectory from a grid of cells and calculated intensities $\beta_{ij}$ obtained with the CF method, we must first observe that the expected intensity in Eq. (1.1) is assumed to depend on the concentrations associated with and residence times of trajectories passing through cell $ij$ but not on the area or shape of the cell.

We can generalize Eq. (1.1) to a set of $m$ distinct cells $G = \{ij_1, ij_2, ..., ij_m\}$, leading to

$$\beta_G = \frac{\sum_{l=1}^{n} t_{Gl} c_l}{\sum_{l=1}^{n} t_{Gl}}, \tag{1.2}$$

where $t_{Gl} = \sum_{ij \text{ in } G} t_{lij}$. Eq. (1.2) can be further transformed

$$\beta_G = \frac{\sum_{ij \in G} \sum_{l=1}^{n} t_{lij} c_l}{\sum_{ij \in G} \sum_{l=1}^{n} t_{ijl}} = \frac{\sum_{ij \in G} t_{ij} \beta_{ij}}{\sum_{ij \in G} t_{ij}},$$

where $t_{ij} = \sum_{l=1}^{n} t_{lij}$ is the total residence time of all trajectories in cell $ij$. Therefore, under the assumptions of the CF method, the natural prediction of pollution concentration for a trajectory is the weighted (according to residence times of the trajectory) average of the intensities of the set of cells that the trajectory passes through.

### 1.2.2 Redistributed Concentration Field

The redistributed CF method is composed of two parts. First, the initial concentration field $\beta_{ij}^{(0)}$ is computed as in Eq. (1.1). And second, the concentrations are iteratively redistributed along trajectories until a stopping criterion is met.

To facilitate the redistribution, each trajectory $l$ is segmented into $n_l$ segments. In this thesis, we based the segmentation on the cells the trajectory passes through. Let $B_{lk}$ represent the concentration field value of the

$k-$th cell trajectory $l$ passes through (that is, $B_{lk}$ corresponds to some cell's concentration field value $\beta_{ij}$).

The concentration associated with the $l$-th trajectory is then partitioned, with each segment $lk$ getting a share proportional to it's concentration field value

$$c_{lk} = c_l \frac{B_{lk}}{\sum_{k=1}^{n_l} B_{lk}}.$$

To complete the iteration, the new concentration field is computed

$$\beta_{ij}^{(t+1)} = \frac{1}{\sum_{l=1}^{n} \sum_{k=1}^{n_l} t_{lkij}} \sum_{l=1}^{n} \sum_{k=1}^{n_l} c_{kl} t_{lkij},$$

where $t_{lkij}$ is the residence time of trajectory $l$'s $k-$th segment in cell $ij$, which is in our case equivalent to $t_{lij}$ if trajectory $l$ resides in cell $ij$ and 0 otherwise. The process is repeated for several iterations. The standard stopping criterion is based on the maximum change, relative to the maximum initial cell value $\frac{\max |\beta_{ij}^{(t+1)} - \beta_{ij}^{(t)}|}{\max \beta_{ij}^{(0)}} < \tau$, for some pre-specified threshold $\tau > 0$. Similar to related work, we set $\tau = 0.01$.

# Chapter 2

# A machine learning-based approach

Our hypothesis was that we could use machine learning to improve on existing methods for pollution prediction and source attribution. In this Chapter we present our methods.

## 2.1 Grid and XYZ data transformation

CF and RCF both work on trajectory data transformed into a grid. If we assume one data sample is one day we end up with one $n \times n$ grid and one pollution value for that day, a total of $n \times n + 1$ values. This transformation discards some data which may be useful (distance and time of the trajectory from station, individual trajectories). To solve that we simply used the most raw trajectory data: a sequential list of trajectory vertices, where each vertex contains the longitude, latitude and height of that point in 3D space. We call this the XYZ transformation. In this case one data sample contains $k$ trajectories and each trajectory has 57 vertices, which gives us a total of $k \times 57 \times 3 + 1$ values for one data sample.

## 2.2 Machine learning models

We included random forest (RF) and bayesian lasso regression (blasso). The RF model could be very good as it can learn to subdivide space based on polluted regions and is robust to overfitting. The blasso model is L1-regularized linear regression (also known as the Bayesian lasso), which penalises regression coefficients the larger their absolute value is. It is also very robust to overfitting. The regularization parameter is treated as a hyper-parameter, so there is no need for tuning [21]. We also tried linear regression but it overfit the data, so the results were not included. We did not use models such as neural networks and support vector machines because they are prone to overfitting and require fine tuning of parameters, which would be very computationally intensive in our case. The RF model uses Breiman's random forests algorithm [9] as implemented in randomForest from the randomForests package

## 2.3 Model explanation

We needed a way to explain why the machine learning models predict as they do. For grid based transformation this would be easily done for regression models just by reading the coefficients, but this would not work for random forest or XYZ transformation.

To solve that we used a black box explanation method presented in [36, 37]. For both grid and XYZ transformation we end up with two $n \times n$ grids: mean and variance. The mean grid tells us how much pollution a certain cell is emitting or absorbing while the variance grid tells us how much the cell values changes from day to day.

### 2.3.1 Grid transformation explanation

We pick a random day, which means that we have $n \times n$ attributes. We then pick one cell for which we want to compute importance and modify

this grid by setting the selected cell's value to 0. We then compare the difference between the results of prediction using the non-modified grid and the modified grid. As this is a sampling method, we must repeat the process above many times for each cell to get reliable results (each sample represents one day of data). Note that when using the blasso model we can also use the regression coefficients directly, as coefficients map one-to-one with grid cells. We present both methods in this work, where we label the direction coefficient method as "blasso direct" and the model explanation method as "blasso explanation".

### 2.3.2 XYZ transformation explanation

For this method we first create an artificial grid. We then sample random trajectories and override some trajectory vertices' longitude and latitude coordinates. We can override a single vertex or many. We then compare the difference of the prediction using the modified trajectories and non-modified trajectories and this serves as importance. As this is a sampling method we must repeat the process above many times for each cell to get reliable results (each sample represents one day of data).

### 2.3.3 Multi-station model

Our goal was to further improve prediction results by using the information from multiple stations at the same time. It turned out this was more complex as we thought initially and the full solution would require additional research that is out of the scope of this thesis. We implemented a simplified multi-station model, which averages grids of multiple stations from the single-station explanation method. We rank the cell pollution values from most to least polluted for every individual station and then compute the average rank across all stations. We can use less ranks than there are cells because the non-extreme rank values will move around from cell to cell a lot due to noisy results.

## 2.4 Visualization

### 2.4.1 Standard approach and scale problem

We use the mean grid from the black box model for visualization. A problem we face is that of great cell value disparity as this causes the color scale of the visualization to stretch, which ruins the visualization. Using a log scale was not enough. We partially solved this by removing the most extreme cell values progressively by using certain thresholds. For example, removing the cells whose values belong in the top and bottom 5%. This gives us multiple images of the same mean grid with increasing number of removed cells. The cells we remove can be either marked with a special symbol to show that they were removed or their color can be clamped to its extreme range color (for example fully white or black). See Figure 2.1 b) and c).

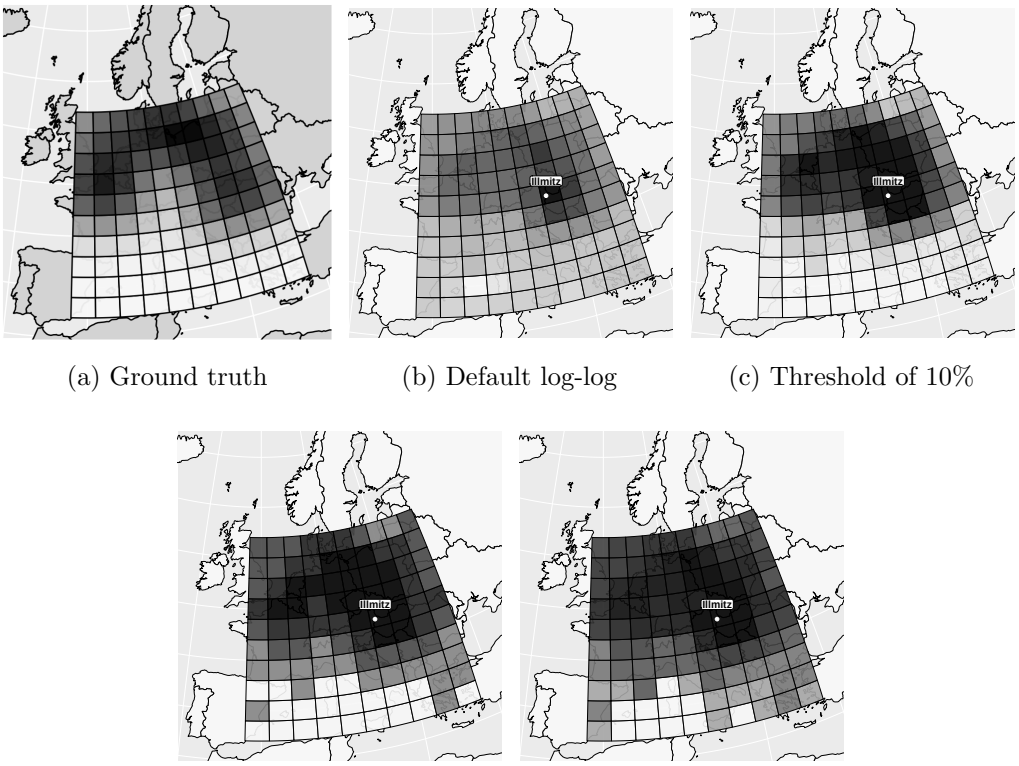
### 2.4.2 Ranking

An alternative approach to removing the extreme-valued cells is to rank the grid cells based on their pollution level. So, for example, if we have 100 cells we get 100 ranks. This is still problematic due to noise as the non-extreme ranks would move a lot from cell to cell. We can improve this method by reducing the number of ranks relative to the number of cells, so if we have 100 cells we use just 10 ranks. This makes the visualization more robust. The downside of this ranked approach is that we lose any kind of physical units. See Figure 2.1 d) and e).

## 2.5 Visualizing various methods

CF and RCF are easily visualized since we can just use their corresponding matrix, where each value in the matrix represents the pollution value for the cell to which it belongs. The grid and XYZ methods are visualized by using the algorithm presented in Section 2.3 and then visualizing the mean grid

computed by that method.




(a) Ground truth (b) Default log-log (c) Threshold of 10%

(d) 5 ranks (e) 10 ranks

Figure 2.1: Examples of different visualization methods results on artificial data using our grid method with random forest model on Illmitz trajectories. Darker cells represent more polluted regions.

# Chapter 3

# Empirical evaluation

This Chapter is divided into 5 parts. First we describe the data, then the evaluation and performance procedure. We then explain our evaluation procedure and performance evaluation. This is followed by a brief overview of the implementation. We conclude this chapter by specifying the model parameters used.

## 3.1 Data

We compiled a data set of pollutant measurements and trajectories for 5 stations: Illmitz, Svratouch, Zingst, Iskrba and Westerland. We have between 10 and 15 years of trajectory and pollutant data, this varies from station to station and also between pollutant types. For pollutants, we have 5 different types of measurements: $PM_{10}$, $PM_{2.5}$, $SO_2$, $NO_2$ and $O_3$,measured in $\mu g/m^3$, but note that not all stations have all of these. Pollutant measurements were taken every 8 hours, starting at 8:00 CET. Pollutant data were obtained from the European Monitoring and Evaluation Programme (EMEP) for monitoring trans-boundary air pollution.

The log function was applied to pollution values because they are non-negative and tend to have a long-tailed distribution. This transforms them to a normal-like distribution as some models are known to perform better

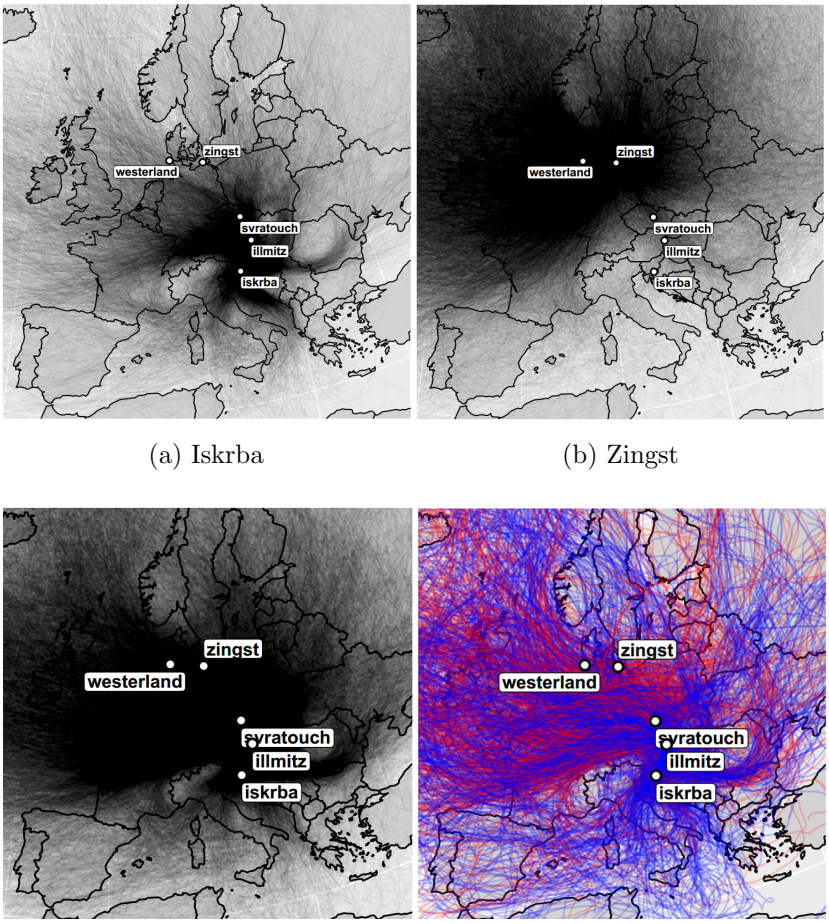
with normally distributed data. Also, MSE would be misleading with a highly skewed distribution. To prevent negative pollution values we added a bias of 1 to all pollution values: $log(p + 1)$. Using this bias does not change relative results between various methods and models, neither does it affect source attribution visualization tests. The RCF methods does not work with negative pollution values (never reaches threshold in certain cases) and negative values do not make sense in this context anyway.

Backward air trajectories were computed by the Norwegian Institute for Air Research (NILU) using the FLEXTRA model. These are 3-D, 7-days backward trajectories with 3-h intervals, computed four times a day ending at 00:00, 06:00, 12:00, and 18:00 UTC. The meteorological data originates from the European Centre for Medium Range Weather Forcasts (ECMWF). Their spatial resolution of is T106 (1.125 x 1.125 degrees), which is relatively high, but not the highest resolution - in several studies, that focus on a single station, higher resolution is used. We include in our analysis only the lowest trajectories that arrive to stations at a few meters above ground, because at this arrival height are expected to have the largest influence. Trajectories also change from year to year. You can see an example of trajectories in Figure 3.1.

## 3.2 Evaluation procedure

The simplest possible model is one which always naively predicts the mean pollution. This approach does not require trajectories (and therefore any transformation) and is used as a baseline for determining the usefulness of a method.

To validate our daily prediction results we use cross-validation (CV). Folds contain either daily or yearly groupings. We choose 5-fold daily CV, where fold indices for each day is the repeating sequence 1 to 5 for all days sorted by their date: $1, 2, 3, 4, 5, 1, 2, 3, ...$ The reason for this is long computation time. We used a per-year CV method for a subset of the data. This




(a) Iskrba (b) Zingst

(c) Illmitz (d) Illmitz trajectories for year 2000 (blue) and 2012 (red).

Figure 3.1: Trajectories visualizations (every 6-th trajectory for clarity).

gave us 15 folds, one fold for each year. The final results were better for that subset, but the relative performance of the $R^2$ metric between methods remained the same. For source attribution visualization, we trained the models on all available data.

## 3.3 Performance evaluation

We used the mean squared error and the coefficient of determination (or $R^2$) metrics:

$$R^2 = 1 - \frac{\text{MSE}_{model}}{\text{MSE}_{baseline}}, \tag{3.1}$$

$R^2$ of 1 means the model is perfect, 0 means it performs equal to baseline and negative values mean it performs worse than baseline.

## 3.4 Implementation overview

To download trajectories we created a Python script which crawls and downloads all trajectory files to disk. A script in the R programming language than further pre-processes trajectory data. The main part of the project is implemented in R. We use the 'caret' package and an optimized version of carets' blasso implementation, based on collaboration with Erik Štrumbelj's ongoing research project. We can describe our project as being made out of 4 parts: 1) data preprocessor, which merges pollution and trajectory data and cleans it, 2) the main data learn/predict part, 3) result processor for cross validation results and 4) visualization generator.

## 3.5 Model parameters

We use 500 trees for random forest. We tried increasing it to 1000 but it increases the computation time too much so we used 500. The caret package 'mtry' parameter for random forest was 200. We also tried different grid sizes: $10 \times 10$, $20 \times 20$ and $30 \times 30$. For reasons discussed in Chapter 4 we present results for $10 \times 10$ grid.

# Chapter 4

# Results

In this chapter we present the results. For each approach we made several test cases to see how prediction works on artificial data. We then show the actual results on real data.

The XYZ method gave the best results in terms of prediction performance but it did not give sensible results when visualized therefore we did not include it in visualizations except to show how it looks. See Figure 6.6 for results. Usually it displays a vertical or horizontal line.

## 4.1 Testing the model explanation methods with artificial data

We needed to test all model explanation methods to see how well they perform on simple and more complex test cases. By creating artificial pollution sources, we could then judge how well these methods work since we know what the results would need to look like. These tests then serve as a benchmark for how much we can believe the visualized pollution regions of all methods when using real data.

We have 3 tests cases. The first is a simple single-point pollution emitter test (see Figure 4.1 a)). We pick a single point on the map with a longitude and latitude radius of 1. All trajectories that pass through this area are

given a certain amount of pollution based on how long they stay in this area. This is the simplest test. For the second test (see Figure 6.1 a)) we also use single-point pollution emitter as for test 1, but we create 4 of these points instead of one. For the third test case we create a more realistic cloudy and noisy pollution source region and see how well the models can predict it (see Figure 6.2 a)).

Here we present test results using the Iskrba station trajectories. We evaluated the results of test method visualization merely by visual inspection.

### 4.1.1 Single-point pollution emitter test

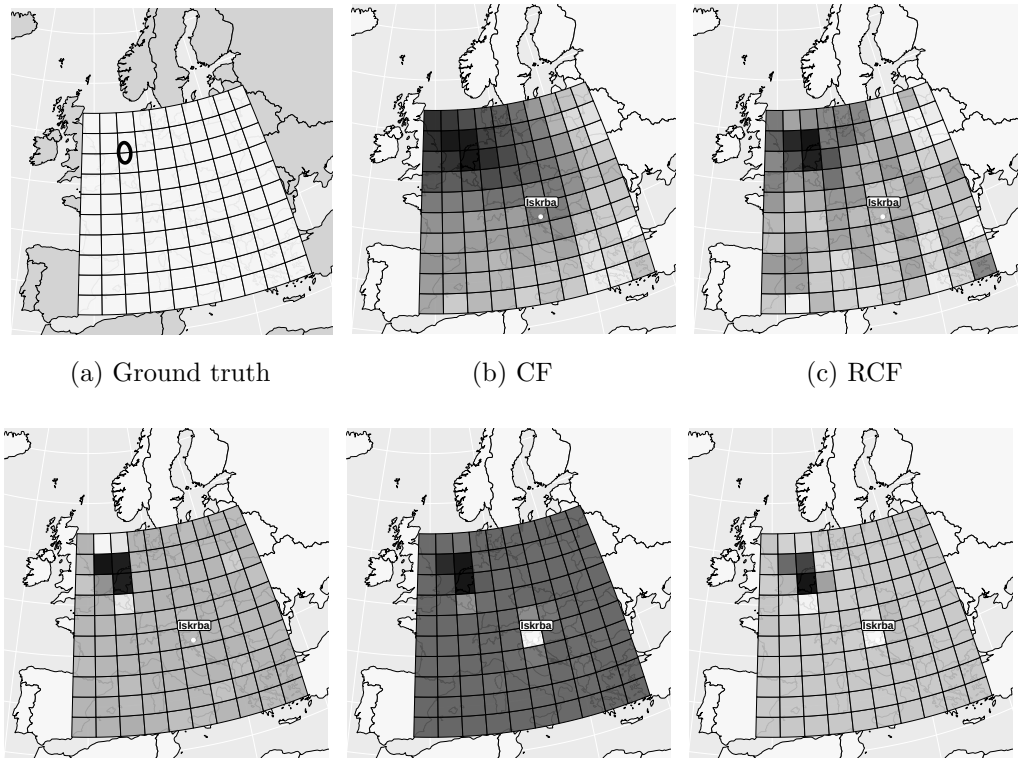
All models perform well on this most simple test. We can see the trailing effect of the CF method in Figure 4.1 b). We can see that RCF fixes the trailing effect in 4.1 c). Both CF and RCF have some noise in other cells while our new methods, grid-blasso and grid-RF don't. Our model explanation method does, however, incorrectly show that the cell with the Iskrba station is very unpolluted.

### 4.1.2 4-point pollution emitter test

The 4-point test in Figure 6.1 shows that our methods can potentially identify multiple point source regions very accurately as there is no noise in the visualization. CF and RCF on the other hand identify blurry regions around the points. We can see that RCF is able to identify regions more sharply than CF. The south-west point is less pronounced in some cases because very few trajectories pass over it. Yet again our model explanation method incorrectly identifies the Iskrba cell as very unpolluted.

### 4.1.3 Single-station cloudy pollution emitter dataset test

We used a less localized, cloudy shape for this test. Real pollution data is most likely not localized to such extremes as in the single-point and 4-point




(a) Ground truth (b) CF (c) RCF

(d) Grid Blasso Direct (e) Grid Blasso Explanation (f) Grid RF

Figure 4.1: Single point test on Iskrba trajectories with log-log scale of physical units.

test, so because of that we consider this test as the most important one. In Figure 6.2 we already come across the problem with visualization scale as it ruins the visualization due to our methods assigning a very large pollution value to the cell in which the Iskrba station resides. We also present this test using a 10-rank visualization in Figure 6.3. We can see that even though CF is very blurry, the underlying shape can be seen. RCF produces a noisier image. Both grid-blasso explanation and grid-RF overestimate pollution in the Iskrba station region. Both grid-blasso visualizations produce a very noisy result. The grid-blasso explanation method overestimated pollution in the Iskrba station region and underestimated it in the actual north-east region, where the most polluted area is. From this we can see that grid-

blasso direct may be a better choice between the two. This test may be problematic due to the fact that not many Iskrba station trajectories pass the most polluted area. This can be seen in Figure 3.1 a).

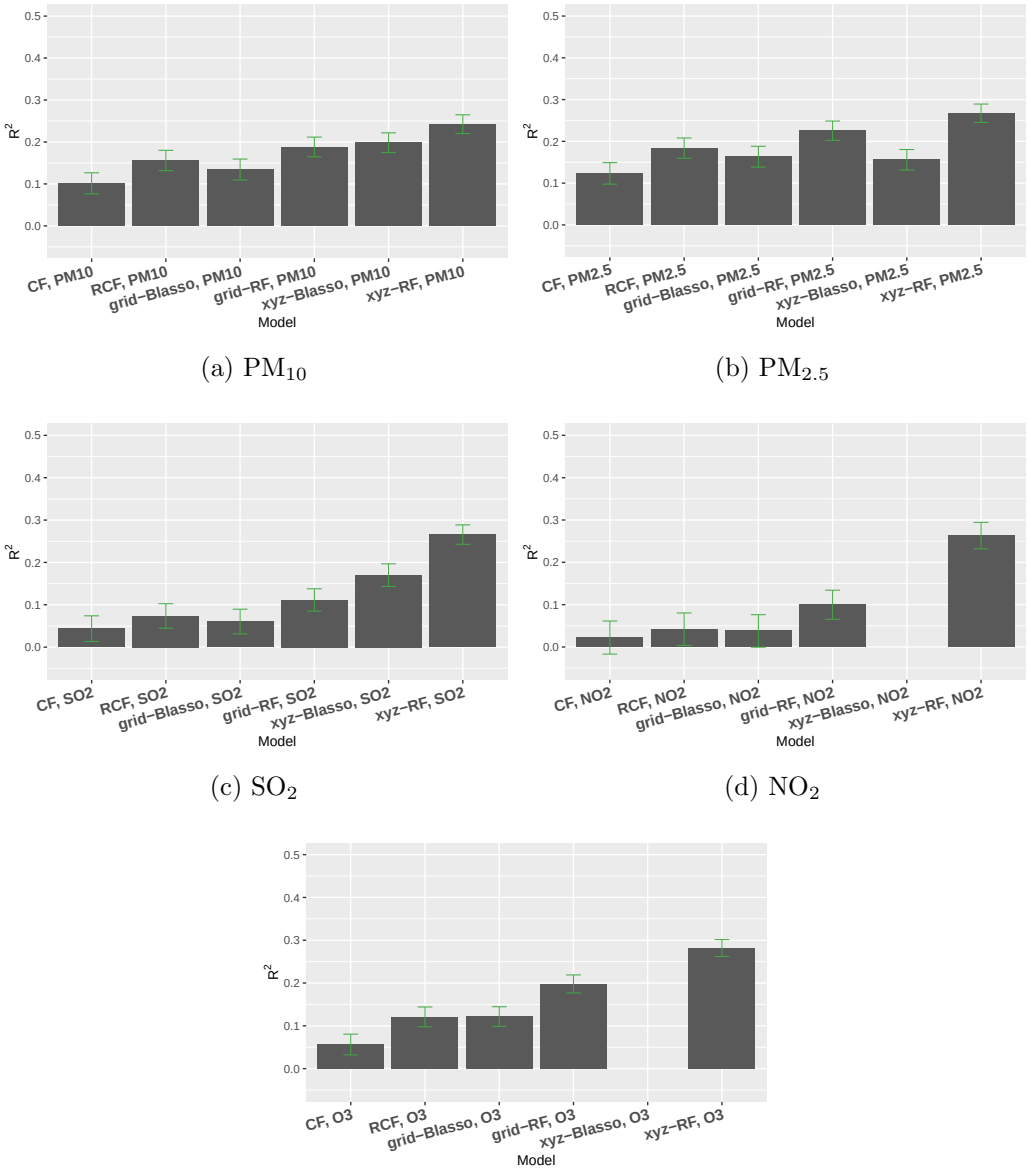
Because of that we made a second cloudy test for the Illmitz station which you can see in Figures 6.4 and 6.5. The results are again not the best, but the general shape is more or less captured by all methods, definitely better than the Iskrba cloud test. The grid-blasso direct method of visualization seems to be better again than the explanation version.

### 4.1.4 Multi-station cloudy pollution emitter dataset test

We used the second cloudy shape to test the multi-station method. For this we computed the test using the cloudy shape test trajectories for every individual station and used that data with our multi-station method. From Figures 6.7 and 6.8 we can see that grid-blasso methods work best, but still a lot worse than we expected.

## 4.2 Pollution prediction results for single-station methods

In this Section we present daily prediction result scores computed on real data and we rank various results. We used data from 5 stations. Prediction results vary between stations. It seems as if there are 2 groups of stations based on prediction performance result: Illimtz, Westerland and Zingst stations explain almost twice the amount of data than Iskrba and Svratouch. See Figure 4.2 for Iskrba station results and Figure 4.3 for Zingst station results. Iskrba belongs to the worst performing group, while Zingst belongs to the better performing group. Plots for $R^2$ also show the standard error as the green interval on the top of each bar.




(a) $PM_{10}$ (b) $PM_{2.5}$

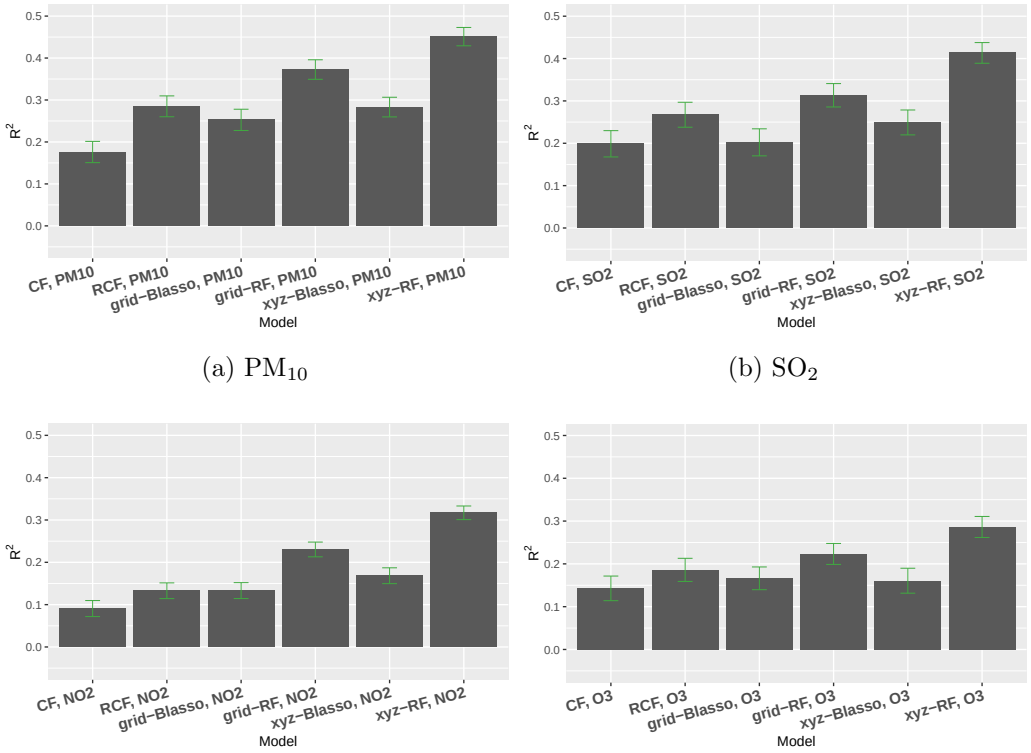
(c) $SO_2$ (d) $NO_2$

(e) $O_3$

Figure 4.2: Iskrba station $R^2$ 5-fold cross validation results. Note that xyz-blasso results for $NO_2$ and $O_3$ are negative and thus not shown.

### 4.2.1 Ranking by pollutants

We wanted to know which pollutants could be best predicted by our models. By ranking pollutants by their mean prediction results we can see which of




(a) $PM_{10}$ (b) $SO_2$

(c) $NO_2$ (d) $O_3$

Figure 4.3: Zingst station $R^2$ 5-fold cross validation results. Note that we are missing Zingst $PM_{2.5}$ data.

them are transported around the world by air mass movements and which are more independent of it. We first computed the mean $R^2$ metric across all methods and model types and stations for each pollutant. These results are presented in Table 4.1 It is known that particles are more easily transported more through air than gasses, which may also explains why $PM_{10}$ and $PM_{2.5}$ have better results. In Tables 4.2, 4.3, 4.4, 4.5, 4.6 we show mean and median pollutant concentration for each station. $O_3$ also stands out as it has the highest concentration of all pollutants, but has the poorest prediction results.

Table 4.1: Mean and median $R^2$ ranking of pollutants across all methods and stations.

| Pollutant | Type | Mean | Median |
|---|---|---|---|
| $PM_{10}$ | Particle | 0.26 | 0.26 |
| $PM_{2.5}$ | Particle | 0.24 | 0.23 |
| $SO_2$ | Gas | 0.20 | 0.19 |
| $NO_2$ | Gas | 0.14 | 0.11 |
| $O_3$ | Gas | 0.12 | 0.16 |

Table 4.2: Mean and median $PM_{10}$ pollutant concentration and $R^2$ values.

| Station | Mean Conc. | Median Conc. | Mean $R^2$ | Median $R^2$ |
|---|---|---|---|---|
| Illmitz | 3.05 | 3.05 | 0.32 | 0.32 |
| Westerland | 2.90 | 2.89 | 0.33 | 0.30 |
| Svratouch | 2.80 | 2.83 | 0.19 | 0.19 |
| Zingst | 2.72 | 2.67 | 0.30 | 0.28 |
| Iskrba | 2.67 | 2.69 | 0.17 | 0.17 |

Table 4.3: Mean and median $PM_{2.5}$ pollutant concentration and $R^2$ values.

| Station | Mean Conc. | Median Conc. | Mean $R^2$ | Median $R^2$ |
|---|---|---|---|---|
| Illmitz | 2.78 | 2.74 | 0.30 | 0.29 |
| Westerland | NA | NA | NA | NA |
| Svratouch | NA | NA | NA | NA |
| Zingst | NA | NA | NA | NA |
| Iskrba | 2.42 | 2.43 | 0.19 | 0.17 |

Table 4.4: Mean and median $NO_2$ pollutant concentration and $R^2$ values.

| Station | Mean Conc. | Median Conc. | Mean $R^2$ | Median $R^2$ |
|---|---|---|---|---|
| Illmitz | 1.23 | 1.17 | 0.13 | 0.11 |
| Westerland | 1.07 | 1.00 | 0.32 | 0.32 |
| Svratouch | 1.14 | 1.06 | 0.04 | 0.05 |
| Zingst | 1.07 | 1.01 | 0.18 | 0.15 |
| Iskrba | 0.42 | 0.36 | 0.03 | 0.04 |

Table 4.5: Mean and median $SO_2$ pollutant concentration and $R^2$ values.

| Station | Mean Conc. | Median Conc. | Mean $R^2$ | Median $R^2$ |
|---|---|---|---|---|
| Illmitz | 0.74 | 0.61 | 0.25 | 0.25 |
| Westerland | 0.45 | 0.35 | 0.17 | 0.17 |
| Svratouch | 0.99 | 0.83 | 0.17 | 0.16 |
| Zingst | 0.53 | 0.41 | 0.27 | 0.26 |
| Iskrba | 0.41 | 0.25 | 0.12 | 0.09 |

Table 4.6: Mean and median $O_3$ pollutant concentration and $R^2$ values.

| Station | Mean Conc. | Median Conc. | Mean $R^2$ | Median $R^2$ |
|---|---|---|---|---|
| Illmitz | 4.03 | 4.16 | 0.11 | 0.11 |
| Westerland | 4.13 | 4.27 | 0.20 | 0.18 |
| Svratouch | 4.15 | 4.19 | 0.10 | 0.08 |
| Zingst | 4.00 | 4.11 | 0.19 | 0.18 |
| Iskrba | 3.95 | 4.02 | 0.01 | 0.12 |

Table 4.7: Mean and median $R^2$ ranking of methods across all pollutants and stations.

| Model | Mean | Median |
|---|---|---|
| CF | 0.12 | 0.12 |
| RCF | 0.17 | 0.17 |
| Grid Blasso | 0.15 | 0.13 |
| Grid RF | 0.23 | 0.22 |
| XYZ Blasso | 0.14 | 0.17 |
| XYZ RF | 0.31 | 0.29 |

### 4.2.2 Ranking by methods and models

Here we present the mean $R^2$ performance of all methods and models. See Table 4.7. We can see that CF performed worst. RCF performed equally well as both grid and XYZ blasso methods. Random forest models performed best, especially the XYZ method which performed almost twice as good as all other methods.

We tried increasing the grid size to see if it can any significant effect. The $20 \times 20$ and $40 \times 40$ grids did not prove to be any better than the $10 \times 10$ grid. There was a very slight increase in performance using a $20 \times 20$ grid, it fell well into the standard error range. The $40 \times 40$ actually showed decreased performance, most likely due to overfitting.

### 4.2.3 Ranking by station performance

Here we present the mean $R^2$ performance and pollution concentration for all stations, see Tables 4.2, 4.3, 4.4, 4.5, 4.6. We see that Iskrba and Svratouch almost always perform much worse than other stations. Iskrba is a backgound station, therefore the pollution at it is less concentrated, which can attribute to more noisy and uncertain results. We can also see that Iskrba indeed has the lowest pollution concentration across all pollutants. We would also expect

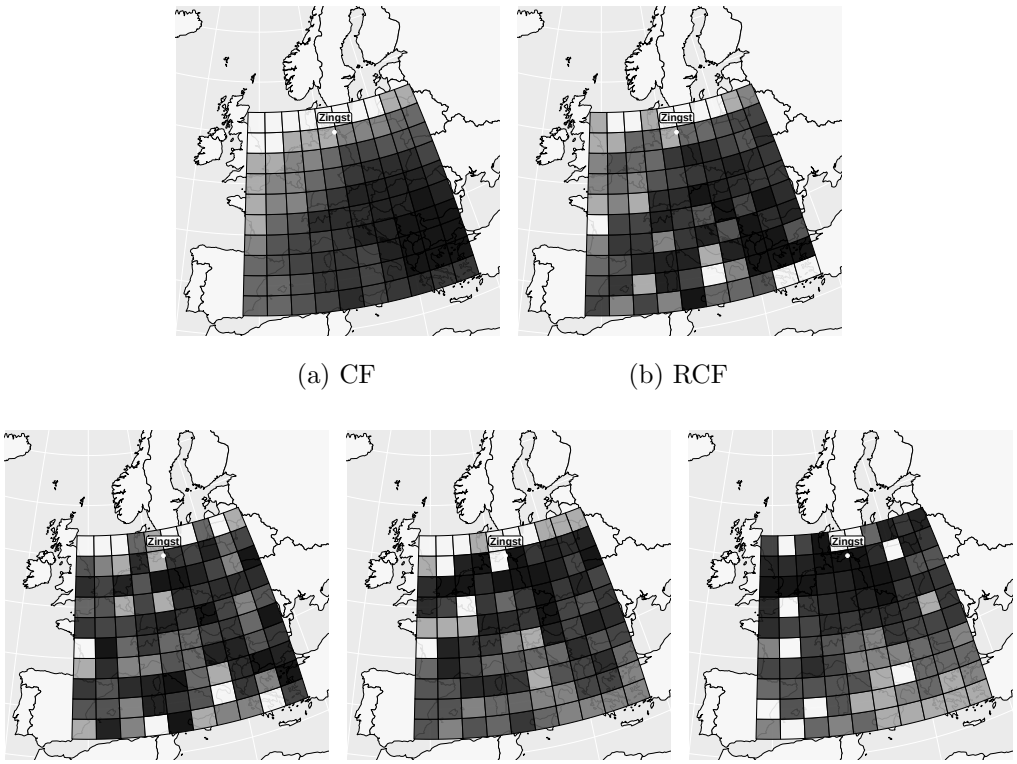
Svratouch to have very low concentrations as it gives very bad prediction results, but that is not true. Westerland $NO_2$ prediction result stands out, as it is much better than others.

## 4.3 Pollution source attribution results for single-station methods

In this Section we present the visualizations of pollution sources as computed by our new and existing models. Note that we were not able to make XYZ method visualization to work. This is a shame because it gives much better prediction performance than other methods. In Section 4.1 we can see that our models are not much better than existing methods in the more complicated cloudy source test. See Figures 6.11 and 4.4 for $PM_{10}$ results, as this pollutant gives best prediction results. It is worth mentioning that the random forest grid method changes quite a bit in some regions.

## 4.4 Pollution source attribution results for multi-station methods

In this Section we present source attribution results for the multistation model presented in Section 2.3.3. The multistation model does not support prediction of pollution because of the way data from multiple station is merged together so we do now have any way of validating its results, for that reason we cannot say how accurate these final results are. See Figures 6.9 and 6.10 for results. We can again that CF and RCF show the most pollution in the south-east grid region. Grid-blasso direct is very noisy so we cannot read any patterns from it. Grid-blass explanation and Grid-RF on the other hand show the same pattern of highest pollution in the middle-top to central region.




(a) CF (b) RCF

(c) Grid Blasso Direct (d) Grid Blasso Explanation (e) Grid RF

Figure 4.4: Zingst $PM_{10}$ source attribution using existing and our new models on real data. Using 10-rank visualization.

## 4.5 Additional results

See Figures 6.12, 6.13, 6.14, 6.15, 6.16, 6.17, 6.18 and 6.19 for additional visualizations.

# Chapter 5

# Discussion and conclusion

The final results of this work were disappointing as we expected our methods to work a lot better. Our methods work very well on the 4-point and single-point tests, however, their performance on the cloud pollution test is questionable. Blasso model is very noisy on these test images which is cause for concern. RF is much less noisy. Increasing the grid size did not notably affect results.

All methods show different source regions for certain pollutant types, see Chapter 6, Figures 6.17 and 6.18 for an example which shows that $SO_2$ comes from the south-eastern region of the grid, while $NO_2$ comes from the south-western region.

Multi-station tests in Section 4.1.4 shows that grid-blasso methods work best, but no method really achieves results that could be said to be better than single station source attribution.

These results presented in Table 4.1 match known data: $O_3$ is not affected by air masses very much, so it is logical that its prediction results are the worst, even though it is the most highly concentrated pollutant. $PM_{10}$ and $PM_{2.5}$ on the other hand are transported by air masses, so our results match domain knowledge. We also noticed that pollutant's $R^2$ prediction results do not increase with that pollutant's concentration levels, which means that some other factors play an important role in prediction performance.

Using the RF model always gives better results than using blasso. Our RF grid and XYZ methods were both better than existing methods, XYZ was almost twice as good as CF and RCF. Unfortunately, we could not properly visualize it. The fact that CF and RCF perform so poorly is alarming, since they are the most commonly used methods in related work. This means that many conclusions in related work which use CF and RCF are questionable.

Overall station prediction results can be placed in two classes, as some stations give twice as good prediction results as others. We cannot make any solid conclusions as to why that is. The only potential reason we can point out is trajectory data, as the physical model used to compute them may give very wrong results in certain unknown, but important regions. Note that our trajectory data was not computed on a high resolution meteorological field. Perhaps results would be better if we were to use trajectories from a higher resolution meteorological field, while using a grid that covers a smaller area, since pollution near the station has a greater effect on daily pollution concentration. Also note that the longer the trajectory is, the larger the probability of it being wrong, as it computed using a physical model. So the further away we go from the station, the larger the probability that we attributed pollution to the wrong source region.

The single station methods in Section 4.3 do show something potentially interesting: in Figure 4.4 our new methods show a very different pollution region that CF and RCF. Our methods show the pollution sources to be near the Zingst station, while CF and RCF show pollution source to be at the south-east of the grid. This may be a very important observation from which we could conclude that CF and RCF are better at showing approximations of actual sources of pollution, while our methods could be better at telling us how more localized air mass movement or terrain profile or other natural or artificial formations near the station affect pollution levels in the air. This may explain why our grid and XYZ methods predict daily pollution better than CF and RCF: pollution far away from the station does not affect the pollution concentration for the current day, while pollution closer to the

station has a much greater effect. Grid-blasso direct and and grid-blasso explanation do not change and both are very noisy compared to grid-RF.

Our results for the Iskrba station match known results: that most pollution comes to Slovenia from the north-east and south regions, we know this from [23].

Because all methods give quite bad results, we cannot conclude if the noise in our methods is really noise or potential regions where the majority of pollution comes from.

If we would have to choose which visualizations to use we would go with CF, RCF and grid-RF. CF and RCF combined could be used to determine far away source regions, grid-RF would be used to determine important local regions. Grid-blasso gives to noisy results to be considered useful.

We cannot make any solid conclusions about the multi-station model in Section 4.4. At most we can say that it works as good as single-station methods.

We were not able to discern the main reason for this poor performance. Reasons could be any of the following:

- Inaccurate trajectory data. Our trajectories were not made on high resolutions fields.
- Extreme seasonal variation due to changing of air mass movements between years, see Figure 3.1 d).
- Changing of the environment (construction of factories, urbanization, deforestation,...).
- Pollutants are more independent of air mass movement than we thought (eg. pollutants mix between air mass currents and stay in the atmosphere longer). This would invalidate all models as they assume pollutants are carried only by air mass currents.
- We do not take in account dispersion and turbulence, though they may be the reason for poor performance, as stated by Scheifinger and Kaiser

[27].

Some of the results of this thesis have been published in [25].

## 5.1 Future work

Using very accurate trajectory data may be crucial in getting better results. Using trajectories from a higher resolution meteorological field combined with a cell grid covering a smaller area may thus give better results. Our models also learn on data from multiple years, which may be problematic due to seasonal variation of the weather, so models that could take that into account may be required. Models that could take in account dispersion and/or turbulent movement of particles could also be tested. Visualization of the XYZ method may shed some new light on pollution sources since it performed twice as good as other methods. A more complex multi-station model could be more resilient to the inherent trajectory inaccuracies.

# Bibliography

[1] Investigation of trajectory statistical methods for locating fugitive emissions sources on a building scale (2010)

[2] Ashbaugh, L.L., Malm, W.C., Sadeh, W.Z.: A residence time probability analysis of sulfur concentrations at Grand Canyon National Park. Atmospheric Environment (1967) 19(8), 1263–1270 (1985)

[3] Brereton, C.A., Johnson, M.R.: Identifying sources of fugitive emissions in industrial facilities using trajectory statistical methods. Atmospheric Environment 51, 46 – 55 (2012)

[4] Dimitriou, K., Kassomenos, P.: Three year study of tropospheric ozone with back trajectories at a metropolitan and a medium scale urban area in greece. Science of The Total Environment 502, 493–501 (2015)

[5] Ferrarese, S.: Sensitivity test of a source-receptor model. Nuovo Cimento-Societe Italiana Di Fisica Sezione C 25(4), 501–511 (2002)

[6] Fleming, Z.L., Monks, P.S., Manning, A.J.: Review: Untangling the influence of air-mass history in interpreting observed atmospheric composition. Atmospheric Research 104, 1–39 (2012)

[7] de Foy, B., Zavala, M., Bei, N., Molina, L.T.: Evaluation of wrf mesoscale simulations and particle trajectory analysis for the milagro field campaign. Atmospheric Chemistry and Physics 9(13), 4419–4438 (2009)

[8] Gilio, A.D., de Gennaro, G., Dambruoso, P., Ventrella, G.: An integrated approach using high time-resolved tools to study the origin of aerosols. Science of The Total Environment 530–531, 28 – 37 (2015)

[9] Gratz, L.E., Keeler, G.J.: Sources of mercury in precipitation to underhill, {VT}. Atmospheric Environment 45(31), 5440 – 5449 (2011)

[10] Gratz, L.E., Keeler, G.J., Morishita, M., Barres, J.A., Dvonch, J.T.: Assessing the emission sources of atmospheric mercury in wet deposition across illinois. Science of The Total Environment 448, 120 – 131 (2013), atmospheric Mercury, Air Pollution, and Associated Effects on Health: A Festschrift to Professor Jerry Keeler.

[11] Hodges, K.I.: Spherical nonparametric estimators applied to the UGAMP model integration for AMIP. Monthly Weather Review 124, 2914–2932 (1996)

[12] Kaiser, A., Scheifinger, H., Spangl, W., Weiss, A., Gilge, S., Fricke, W., Ries, L., Cemas, D., Jesenovec, B.: Transport of nitrogen oxides, carbon monoxide and ozone to the alpine global atmosphere watch stations jungfraujoch (switzerland), zugspitze and hohenpeissenberg (germany), sonnblick (austria) and mt. krvavec (slovenia). Atmospheric Environment 41(40), 9273 – 9287 (2007)

[13] Karaca, F., Camci, F.: Distant source contributions to PM 10 profile evaluated by SOM based cluster analysis of air mass trajectory sets. Atmospheric Environment 44(7), 892–899 (2010)

[14] Keeler, G., Samson, P.: On the spatial representativeness of trace element ratios 268, 115–132 (1989)

[15] Kong, X., He, W., Qin, N., He, Q., Yang, B., Ouyang, H., Wang, Q., Xu, F.: Comparison of transport pathways and potential sources of PM 10 in two cities around a large Chinese lake using the modified trajectory analysis. Atmospheric Research 122, 284–297 (2013)

[16] Lee, S., Ashbaugh, L.: Comparison of multi-receptor and single-receptor trajectory source apportionment (tsa) methods using artificial sources. Atmospheric Environment 41(6), 1119 – 1127 (2007)

[17] Lin, C.H., Wu, Y.L., Chang, K.H., Lai, C.H.: A method for locating influential pollution sources and estimating their contributions. Environmental Modeling & Assessment 9(2), 129–136 (2004)

[18] Makra, L., Ionel, I., Csépe, Z., Matyasovszky, I., Lontis, N., Popescu, F., Sümeghy, Z.: The effect of different transport modes on urban PM 10 levels in two European cities. Science of the Total Environment 458, 36–46 (2013)

[19] Makra, L., Matyasovszky, I., Guba, Z., Karatzas, K., Anttila, P.: Monitoring the long-range transport effects on urban PM10 levels using 3D clusters of backward trajectories. Atmospheric Environment 45(16), 2630–2641 (2011)

[20] Methven, J., Evans, M., Simmonds, P., Spain, G.: Estimating relationships between air mass origin and chemical composition. Journal of Geophysical Research: Atmospheres 106(D5), 5005–5019 (2001)

[21] Park, T., Casella, G.: The bayesian lasso. Journal of the American Statistical Association 103(482), 681–686 (2008)

[22] van Pinxteren, D., Bruggemann, E., Gnauk, T., Muller, K., Thiel, C., Herrmann, H.: A gis based approach to back trajectory analysis for the source apportionment of aerosol constituents and its first application. Journal of Atmospheric Chemistry 67(1), 1–28 (2010)

[23] Poberžnik, M., Štrumbelj, E.: The effects of air mass transport, seasonality, and meteorology on pollutant levels at the iskrba regional background station (1996–2014). Atmospheric Environment 134, 138–146 (2016)

[24] Riuttanen, L., Hulkkonen, M., Maso, M.D., Junninen, H., Kulmala, M.: Trajectory analysis of atmospheric transport of fine particles, SO 2, NO x and O 3 to the SMEAR II station in Finland in 1996–2008. Atmospheric Chemistry and Physics 13(4), 2153–2164 (2013)

[25] Ropret, M., Gašparac, G., Štrumbelj, E.: Pollution source attribution using air mass back-trajectories : a machine learning approach. In: Zajc, T. (ed.) Zbornik petindvajsete mednarodne Elektrotehniške in računalniške konference ERK 2016. pp. 95–98. IEEE Slovenia (2016)

[26] Salvador, P., Artíñano, B., Querol, X., Alastuey, A.: A combined analysis of backward trajectories and aerosol chemistry to characterise long-range transport episodes of particulate matter: the madrid air basin, a case study. Science of the total environment 390(2), 495–506 (2008)

[27] Scheifinger, H., Kaiser, A.: Validation of trajectory statistical methods. Atmospheric Environment 41(39), 8846 – 8856 (2007)

[28] Schichtel, B.A., Gebhart, K.A., Barna, M.G., Malm, W.C.: Association of airmass transport patterns and particulate sulfur concentrations at big bend national park, texas. Atmospheric Environment 40(5), 992 – 1006 (2006)

[29] Seibert, P., Kromp-Kolb, H., Baltensperger, U., Jost, D., Schwikowski, M., Kasper, A., Puxbaum, H.: Trajectory analysis of aerosol measurements at high alpine sites. Transport and Transformation of Pollutants in the Troposphere pp. 689–693 (1994)

[30] Seibert, P.: Inverse modelling of sulfur emissions in europe based on trajectories pp. 147–154 (2013)

[31] Stohl, A., Forster, C., Eckhardt, S., Spichtinger, N., Huntrieser, H., Heland, J., Schlager, H., Wilhelm, S., Arnold, F., Cooper, O.: A backward modeling study of intercontinental pollution transport using air-

craft measurements. Journal of Geophysical Research: Atmospheres 108(D12), n/a–n/a (2003), 4370

[32] Stohl, A., Forster, C., Frank, A., Seibert, P., Wotawa, G.: Technical note: The lagrangian particle dispersion model flexpart version 6.2. Atmospheric Chemistry and Physics 5(9), 2461–2474 (2005)

[33] Stohl, A., Seibert, P., Arduini, J., Eckhardt, S., Fraser, P., Greally, B.R., Lunder, C., Maione, M., Mühle, J., O'Doherty, S., Prinn, R.G., Reimann, S., Saito, T., Schmidbauer, N., Simmonds, P.G., Vollmer, M.K., Weiss, R.F., Yokouchi, Y.: An analytical inversion method for determining regional and global emissions of greenhouse gases: Sensitivity studies and application to halocarbons. Atmospheric Chemistry and Physics 9(5), 1597–1620 (2009)

[34] Stohl, A.: Trajectory statistics-a new method to establish source-receptor relationships of air pollutants and its application to the transport of particulate sulfate in europe. Atmospheric Environment 30(4), 579–587 (1996)

[35] Stohl, A., Eckhardt, S., Forster, C., James, P., Spichtinger, N., Seibert", P.: "A replacement for simple back trajectory calculations in the interpretation of atmospheric trace substance measurements". Atmospheric Environment 36(29), 4635 – 4648 (2002)

[36] Štrumbelj, E., Kononenko, I.: A general method for visualizing and explaining black-box regression models. In: International Conference on Adaptive and Natural Computing Algorithms. pp. 21–30. Springer (2011)

[37] Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. Knowledge and information systems 41(3), 647–665 (2014)

[38] Viana, M., Kuhlbusch, T., Querol, X., Alastuey, A., Harrison, R., Hopke, P., Winiwarter, W., Vallius, M., Szidat, S., Prevot, A., et al.: Source apportionment of particulate matter in europe: a review of methods and results. Journal of Aerosol Science 39(10), 827–849 (2008)

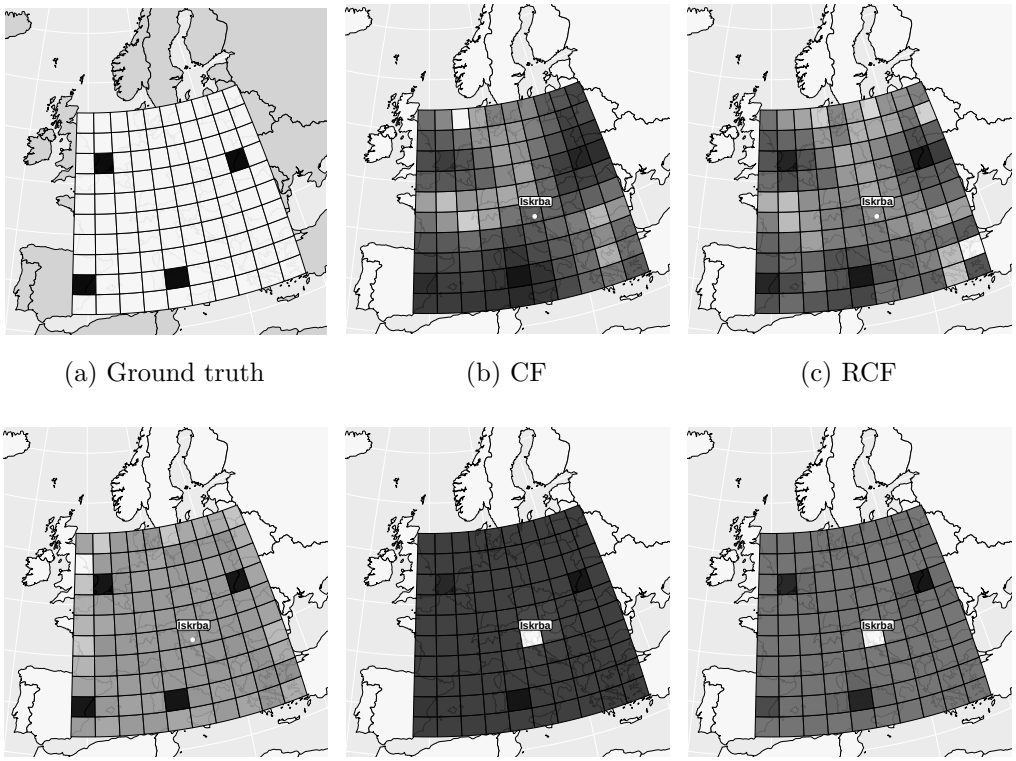
[39] Ying-Kuang Hsua, Thomas M. Holsena, P.K.H.: Comparison of hybrid receptor models to locate PCB sources in Chicago. Atmospheric Environment 37(4), 545–562 (2003)

[40] Zhang, Z.Y., Wong, M.S., Lee, K.H.: Estimation of potential source regions of PM 2.5 in Beijing using backward trajectories. Atmospheric Pollution Research 6(1) (2015)

[41] Zhou, L., Hopke, P.K., Liu, W.: Comparison of two trajectory based models for locating particle sources for two rural new york sites. Atmospheric Environment 38(13), 1955 – 1963 (2004)

# Chapter 6

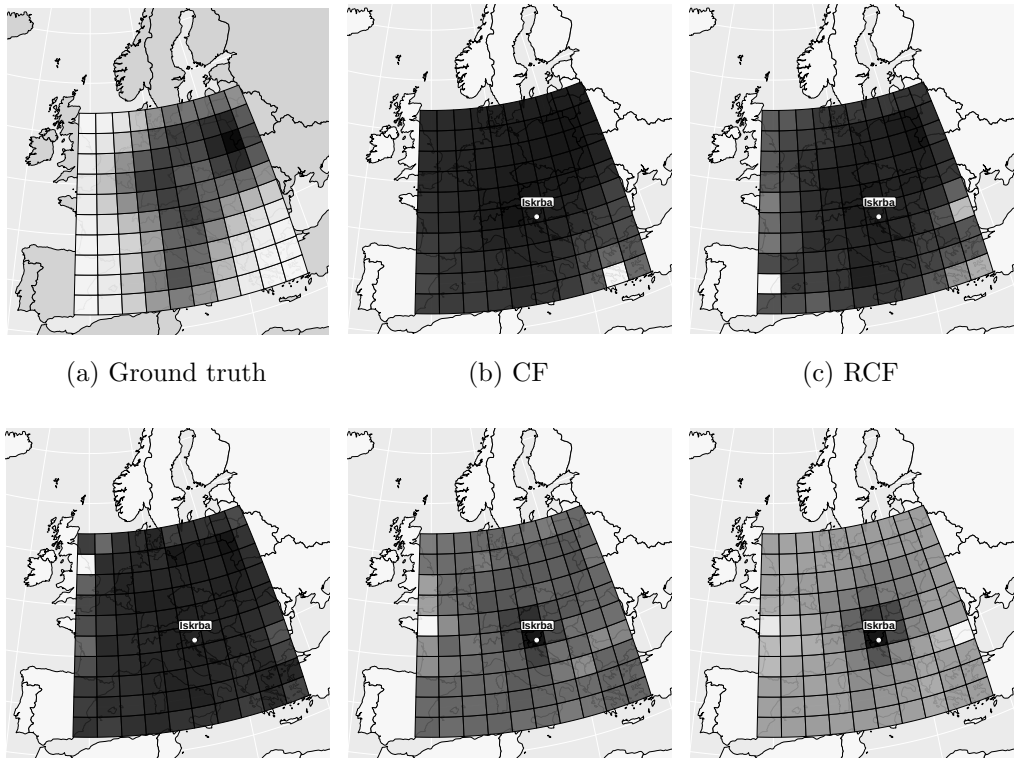
# Appendix




(a) Ground truth (b) CF (c) RCF

(d) Grid Blasso Direct (e) Grid Blasso Explanation (f) Grid RF

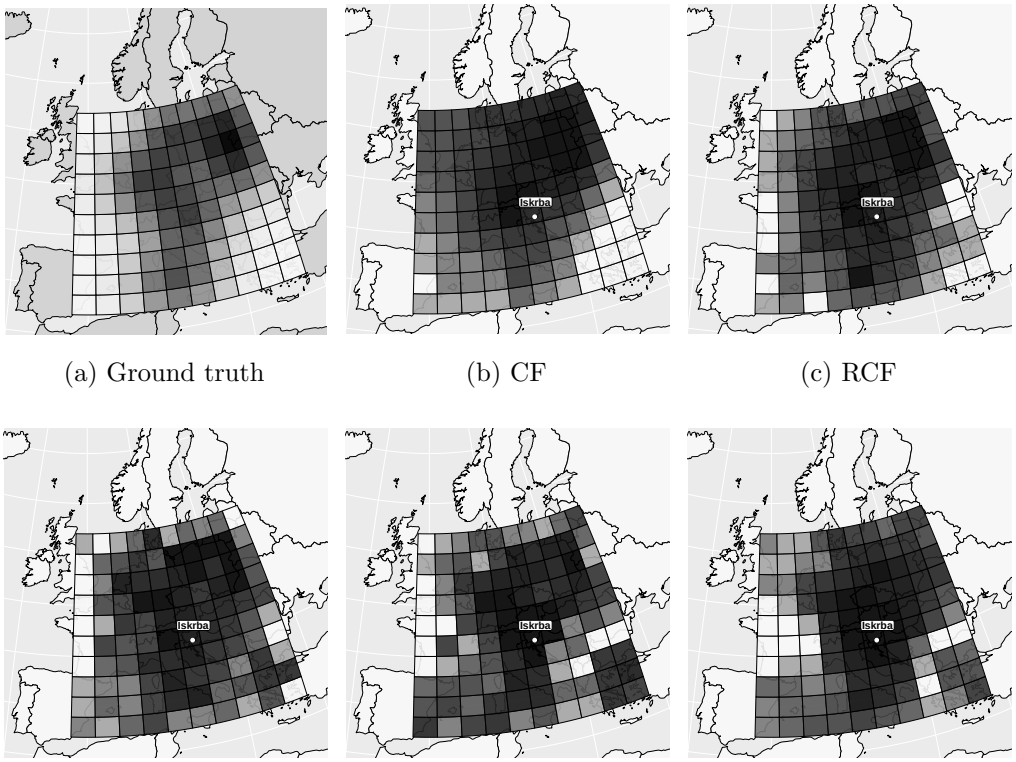
Figure 6.1: The 4 point test on Iskrba trajectories with log-log scale of physical units.




(a) Ground truth (b) CF (c) RCF

(d) Grid Blasso Direct (e) Grid Blasso Explanation (f) Grid RF

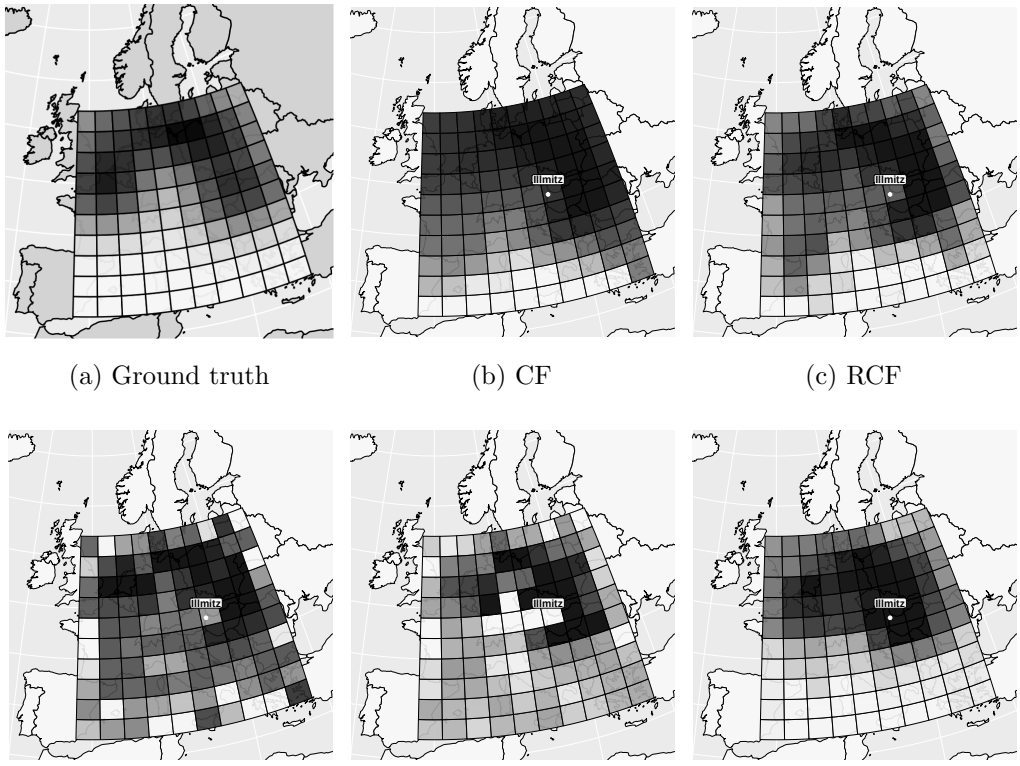
Figure 6.2: Complex cloudy shape test on Iskrba trajectories with log-log scale of physical units.




(a) Ground truth (b) CF (c) RCF

(d) Grid Blasso Direct (e) Grid Blasso Explanation (f) Grid RF

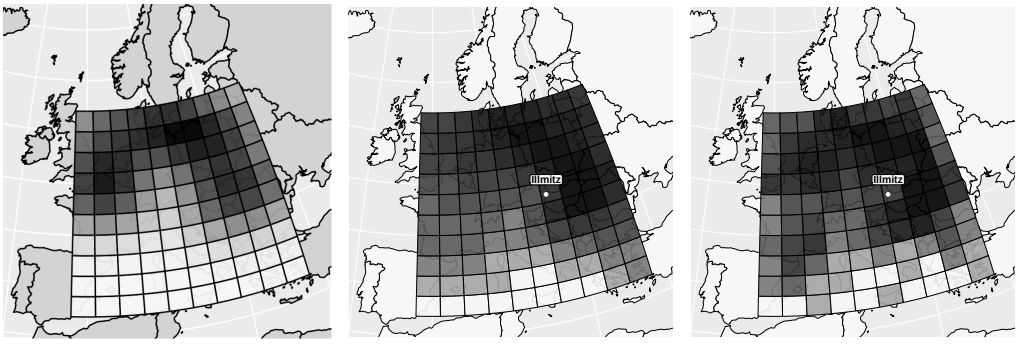
Figure 6.3: Complex cloudy shape test on Iskrba trajectories using 10-class rank visualization.



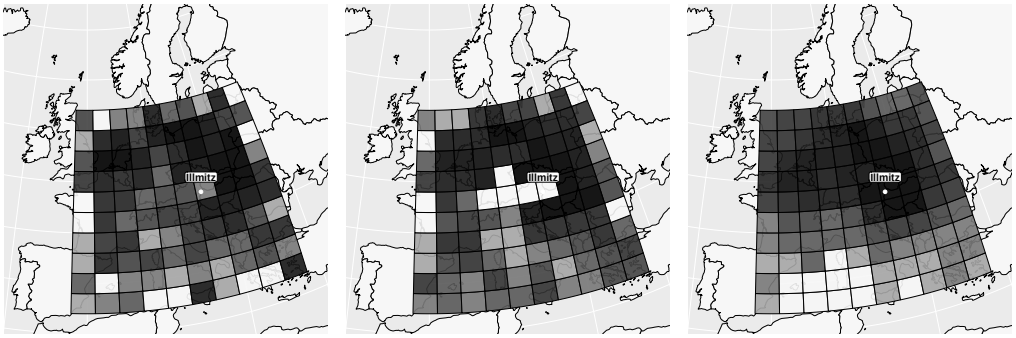

(a) Ground truth (b) CF (c) RCF

(d) Grid Blasso Direct (e) Grid Blasso Explanation (f) Grid RF

Figure 6.4: The second complex cloudy shape test on Illmitz trajectories using log-log scale of physical units and clamping the highest and lowest 10-th percentile values to their extremes, see Section 2.4 for percentile explanation.

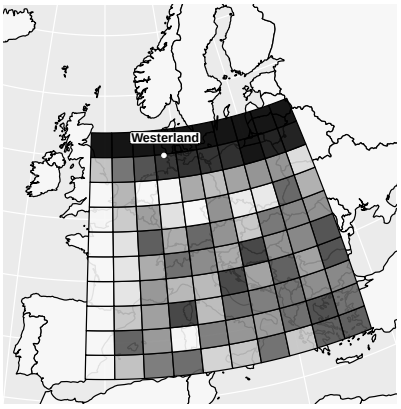


(a) Ground truth (b) CF (c) RCF



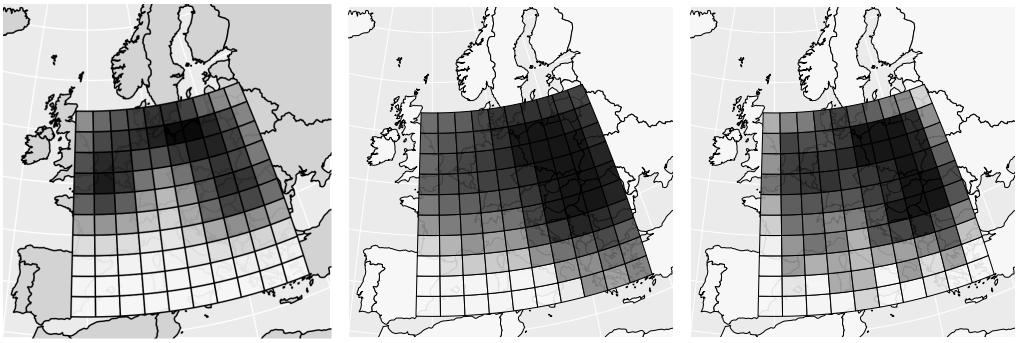
(d) Grid Blasso Direct (e) Grid Blasso Explanation (f) Grid RF

Figure 6.5: The second complex cloudy shape test on Illmitz trajectories using 10-class rank visualization.

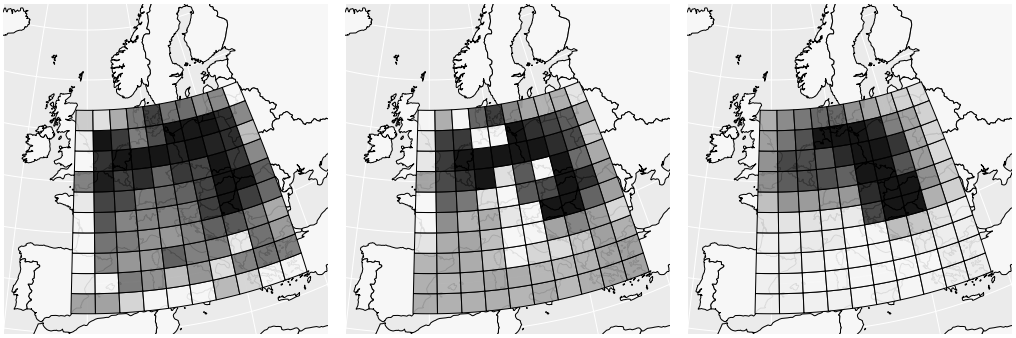



(a) XYZ problem

Figure 6.6: Westerland XYZ method. This is how XYZ visualizations typically look like: a horizontal or vertical line shows up.

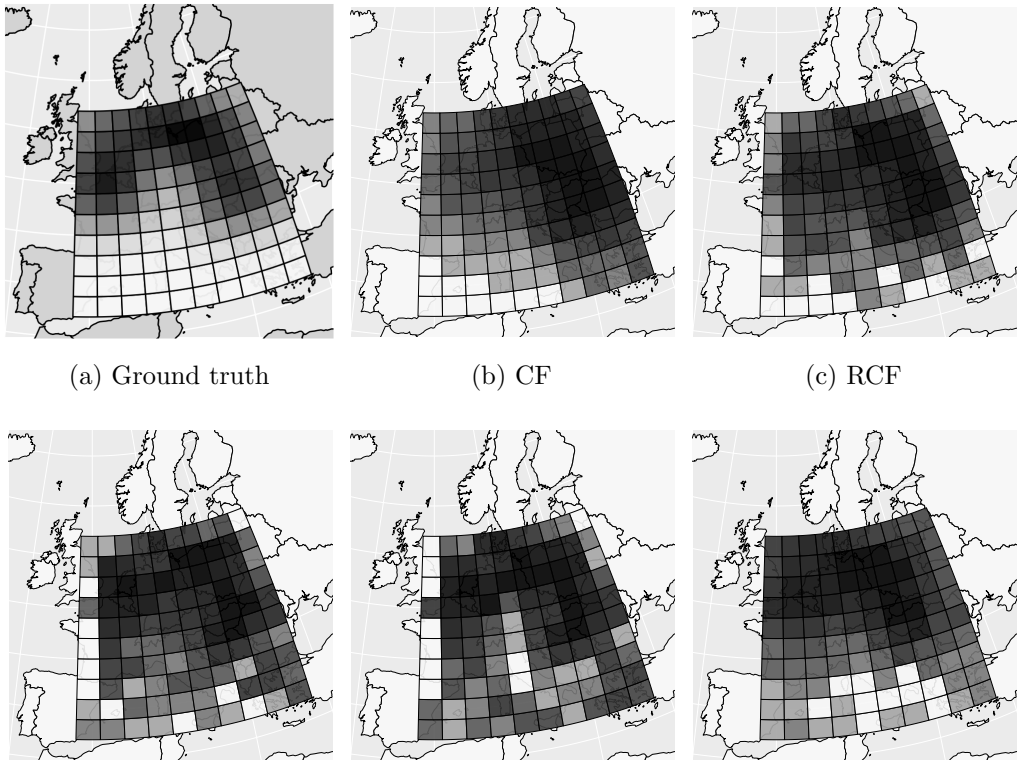


(a) Ground truth (b) CF (c) RCF



(d) Grid blasso direct (e) Grid blasso explanation (f) Grid RF

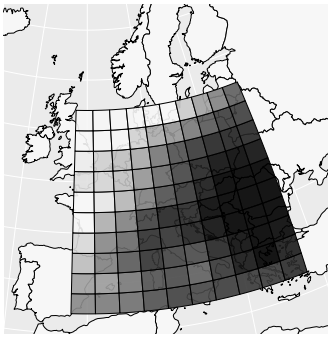
Figure 6.7: Multistation test using log-log scale and clamping the 10-th percentile of the most extreme values.



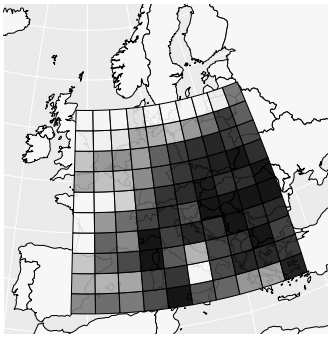
(a) Ground truth (b) CF (c) RCF

(d) Grid blasso direct (e) Grid blasso explanation (f) Grid RF

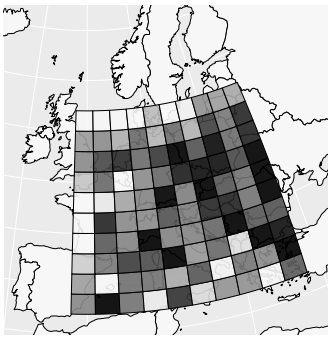
Figure 6.8: Multistation test using 10-rank visualization.



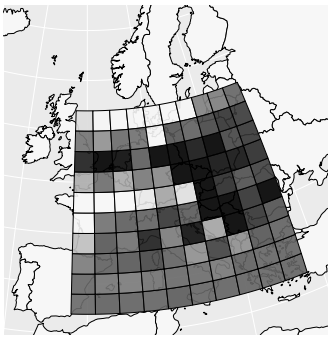
(a) CF



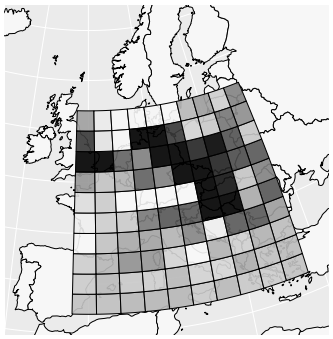
(b) RCF



(c) Grid Blasso Direct

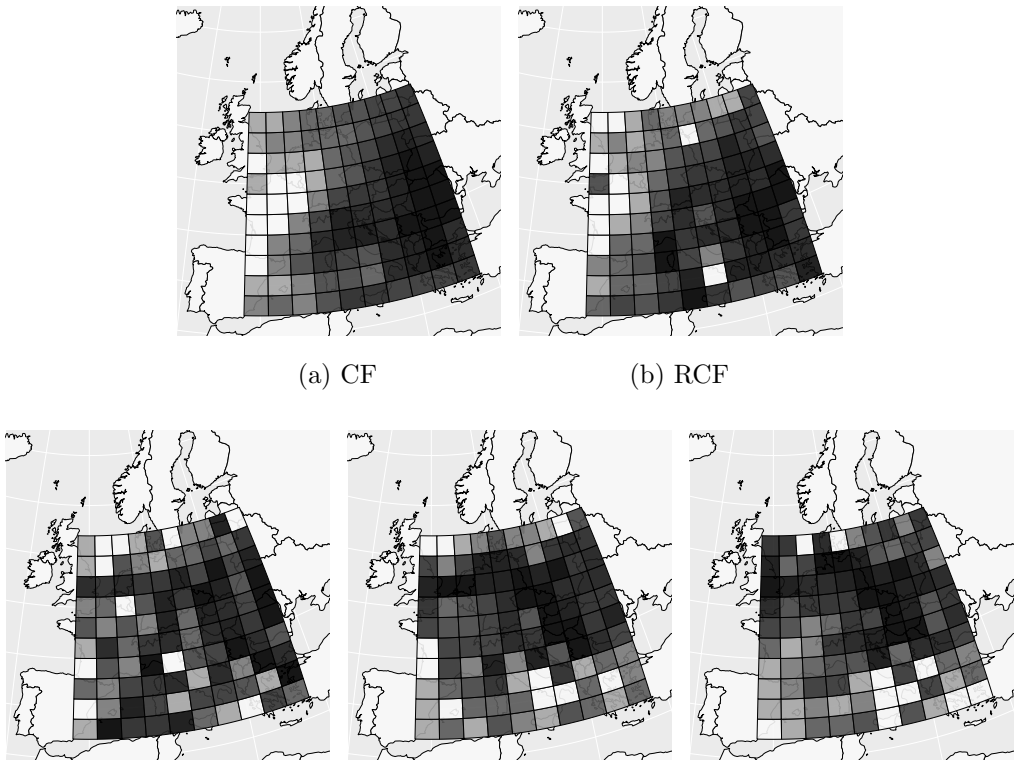


(d) Grid Blasso Explanation



(e) Grid RF

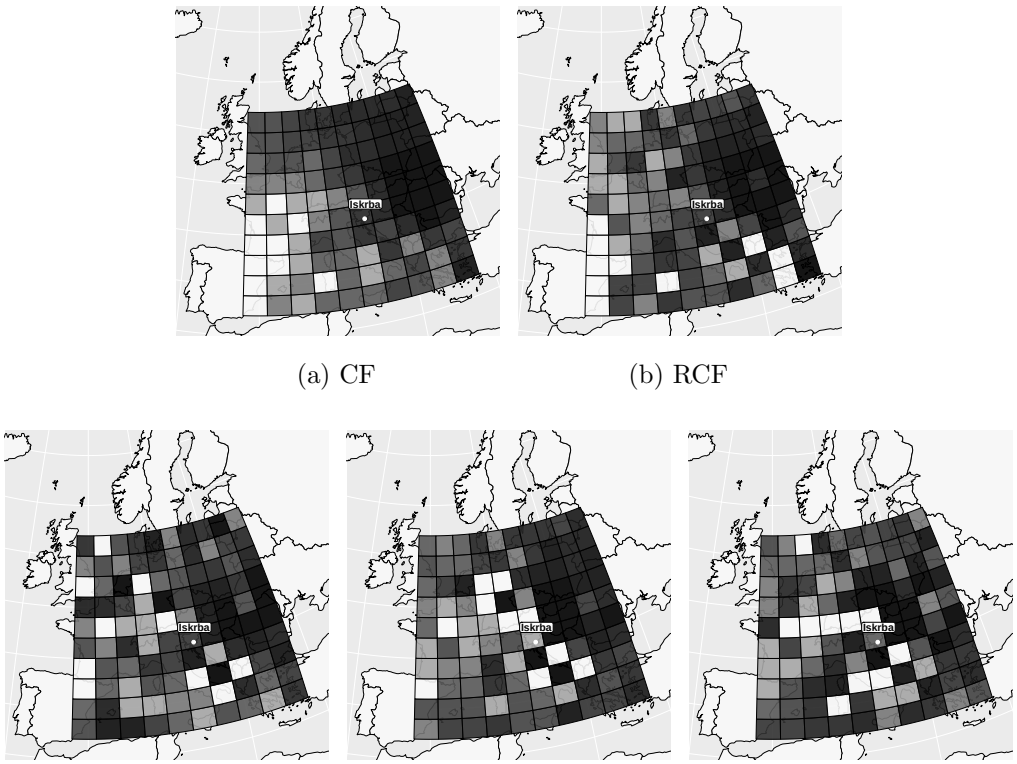
Figure 6.9: Multistation results using log-log scale and clamping the 10-th percentile of the most extreme values.



(a) CF (b) RCF

(c) Grid Blasso Direct (d) Grid Blasso Explanation (e) Grid RF

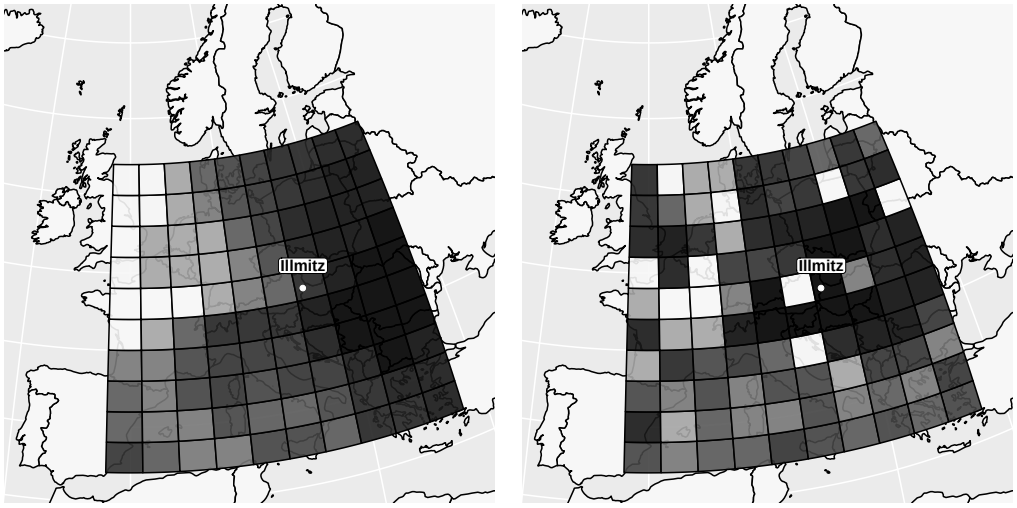
Figure 6.10: Multistation results using 10-rank visualization.




(a) CF (b) RCF

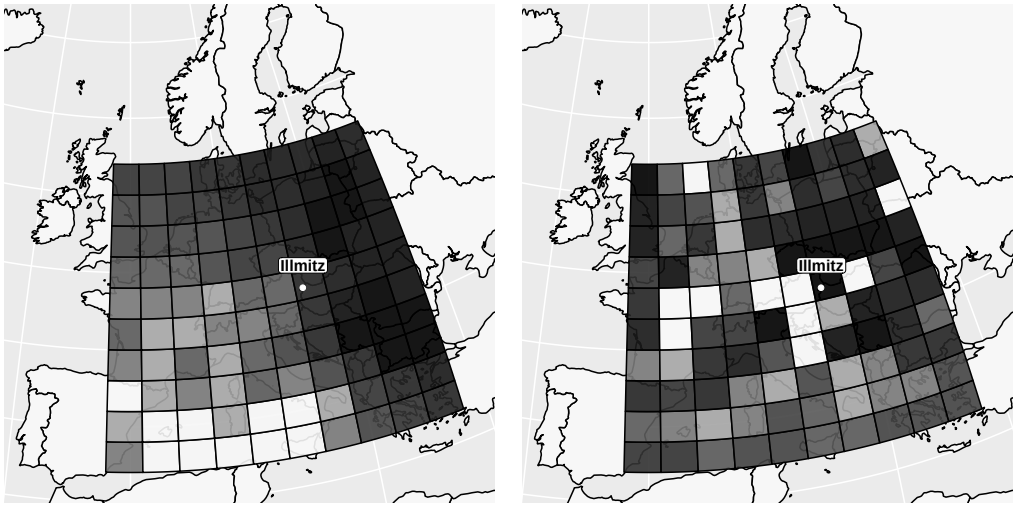
(c) Grid Blasso Direct (d) Grid Blasso Explanation (e) Grid RF

Figure 6.11: Iskrba $PM_{10}$ source attribution using existing and our new models on real data. Using 10-rank visualization.



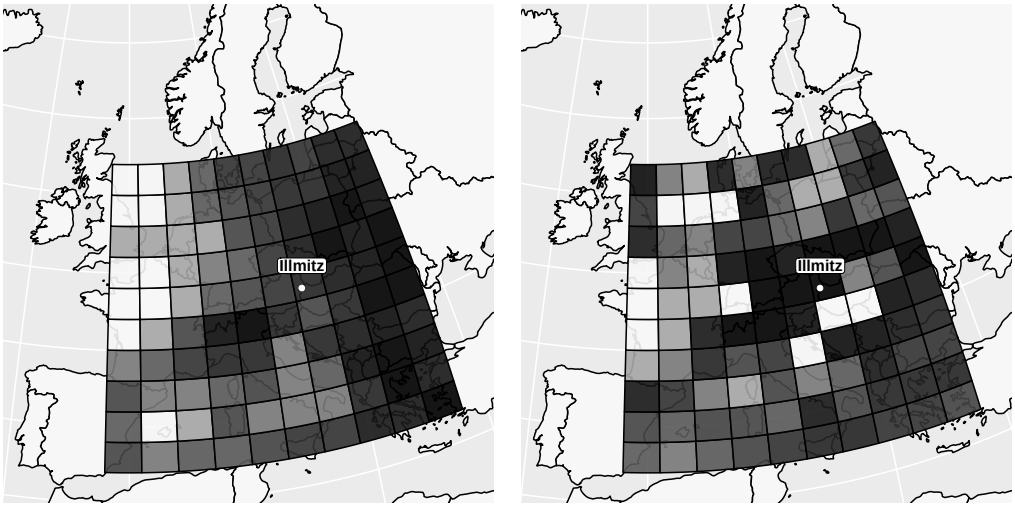

(a) CF (b) Grid RF

Figure 6.12: Illmitz 10-rank single-station $PM_{2.5}$ visualization.



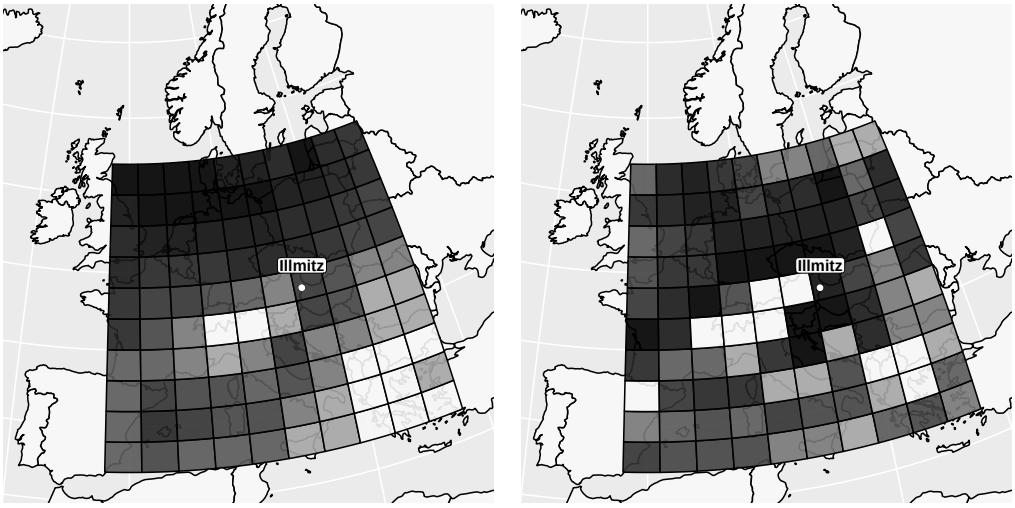

(a) CF (b) Grid RF

Figure 6.13: Illmitz 10-rank single-station $SO_2$ visualization.



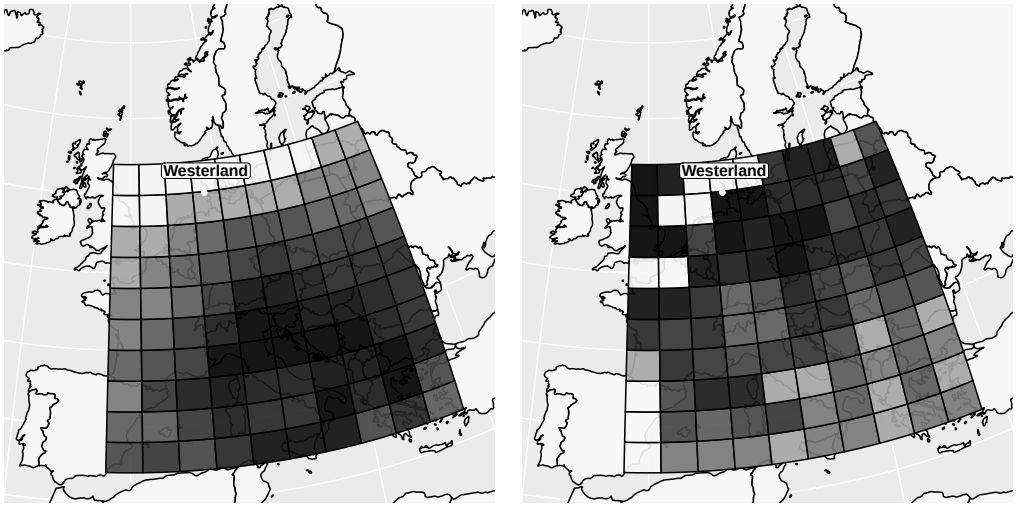

(a) CF (b) Grid RF

Figure 6.14: Illmitz 10-rank single-station $NO_2$ visualization.



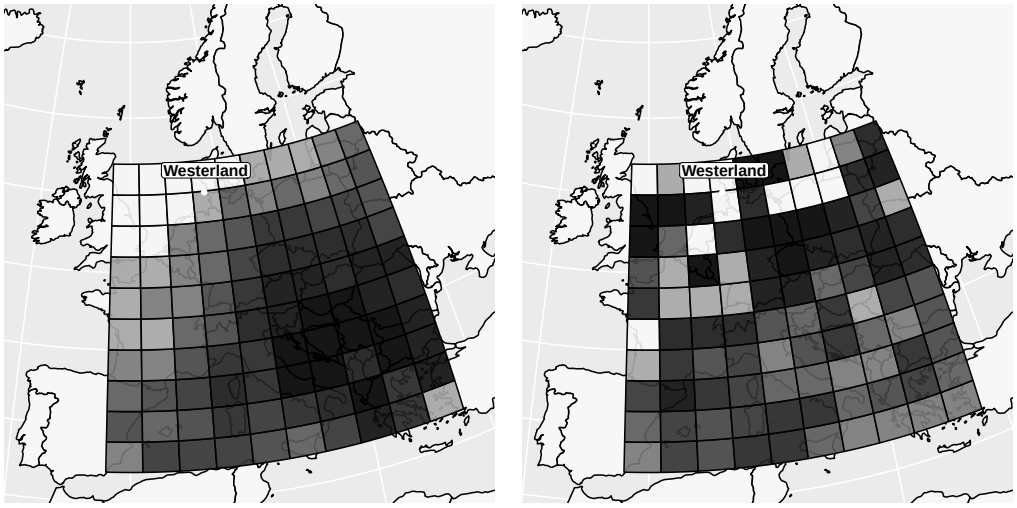

(a) CF (b) Grid RF

Figure 6.15: Illmitz 10-rank single-station $O_3$ visualization.



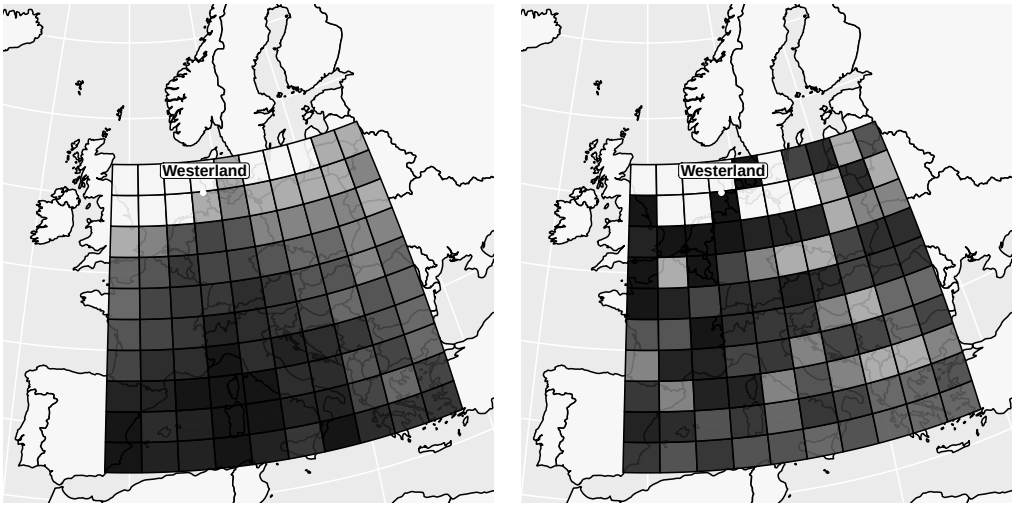

(a) CF (b) Grid RF

Figure 6.16: Westerland 10-rank single-station $PM_{10}$ visualization.



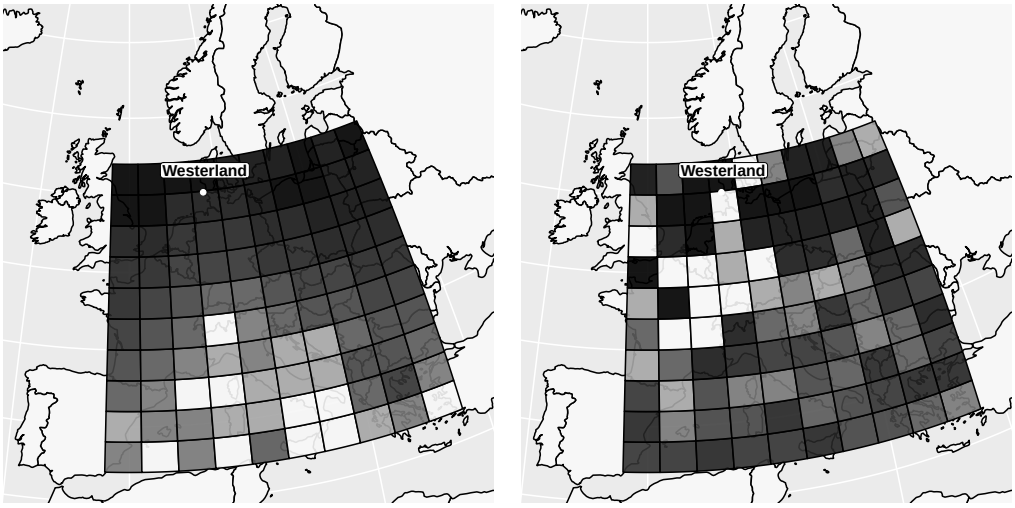

(a) CF (b) Grid RF

Figure 6.17: Westerland 10-rank single-station $SO_2$ visualization.




(a) CF (b) Grid RF

Figure 6.18: Westerland 10-rank single-station $NO_2$ visualization.




(a) CF (b) Grid RF

Figure 6.19: Westerland 10-rank single-station $O_3$ visualization.