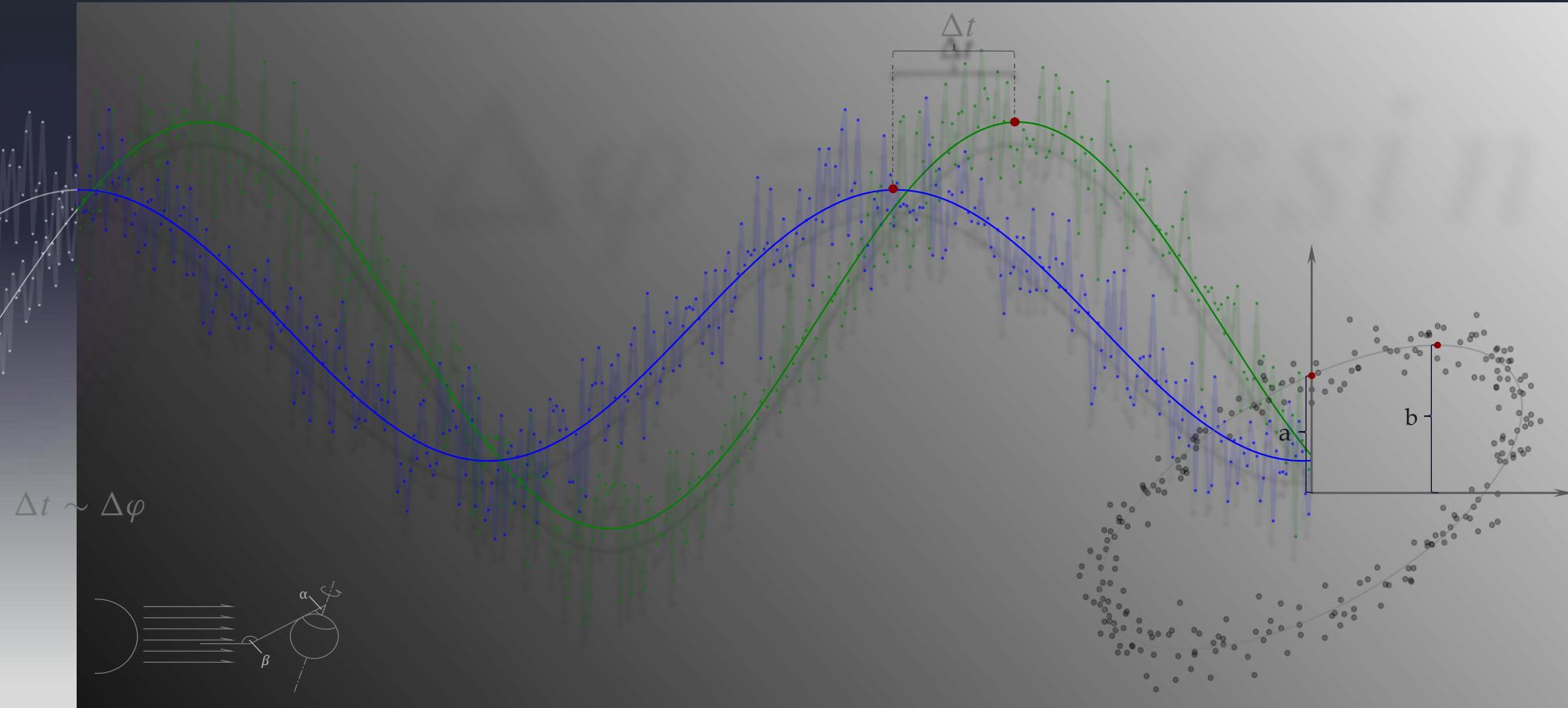


# Datenanalyse auf Wetterdaten

Über verschiedene Analysemethoden auf Wetterparameterzeitreihen  
und einem ungewöhnlichen Wettereffekt

Eine leicht verständliche Zusammenfassung eines naturwissenschaftlichen Datenanalyseprojekts,  
mit etwas Witz und Spaß, manchmal etwas abschweifend  
...und wirklich nur sehr wenigen Gleichungen ☺



Erste Korrekturfassung

1.1

15.05.2021

# *Das kommt dabei raus, wenn ein Naturwissenschaftler nun auch noch Data Science studiert...*

Der Autor über sich:

*Ich habe Geowissenschaften (B.Sc.) an der Universität Potsdam studiert und entschied mich anschließend entgegen einem anfänglich geplanten Master-Studium im geowissenschaftlichen Bereich, doch lieber Data Science an der Hochschule Anhalt zu studieren.*

*Was ich derzeit auch noch tue ...*

*Ich liebe Datenanalyse. Das Finden von Mustern in Daten ist für mich super spannend und fasziniert mich extrem.*

*Im Grunde genommen spielt es fast keine Rolle, was diese Daten eigentlich abbilden. Denn schon während und nach meinem Bachelor-Studium arbeitete ich als HiWi am Leibnitz-Institut für Agrartechnik und Bioökonomie in Potsdam in Bereichen weit außerhalb der Geowissenschaften und analysierte Daten. Zuvor hatte ich am sehr renommierten GeoForschungsZentrum GFZ in Potsdam als HiWi gearbeitet, wo ich mich ausschließlich mit (hydro-)geologischen Daten auseinander setzte. Dort kam ich das erste Mal mit Modellierung und maschinellen Lernen in Berührung und verfasste dort eine Bachelorarbeit, in der ich einen Machine-Learning-Ansatz nutzte, um Messdaten anzupassen.*

*Data Science war also schon im Bachelor-Studium ein großes Thema für mich.*

*Derzeit arbeite ich neben dem Studium bei der Deutschen Telekom IT GmbH als Praktikant in einem Datenanalystenteam und entwickle Zeitreihenanalysebasierte Vorhersagemodelle.*

*Vorher lernte ich übrigens Koch und Fahrzeuglackierer, studierte ein paar Semester „Bauingenieurwesen“ an der FH Potsdam und „Umwelttechnik und Erneuerbare Energien“ an der HTW Berlin und arbeitete hier und da.*

*Ein bewegtes Leben liegt also bereits hinter mir...*

*...und nebenbei bin ich begeisterter Rennradfahrer, Hobbykünstler und Hobbyfotograf, Windsurfer und verbringe viel Zeit mit eigenen kleinen Data Science Projekten.*

Alexander Prinz, Köthen (2021)

Email	: a_prinz@web.de
GitHub	: <a href="https://github.com/FlatEric86/Datenanalyse-auf-Wetterdaten/raw/main/Datenanalyse_auf_Wetterdaten_Alexander_Prinz.pdf">https://github.com/FlatEric86/Datenanalyse-auf-Wetterdaten/raw/main/Datenanalyse_auf_Wetterdaten_Alexander_Prinz.pdf</a>
LinkedIn	: <a href="https://www.linkedin.com/in/alexander-prinz-4644b7184">https://www.linkedin.com/in/alexander-prinz-4644b7184</a>
Kunst und Fotografie	: <a href="https://www.alexander-prinz-art.de">https://www.alexander-prinz-art.de</a>

# Vorwort

Diese Arbeit ist keine rein wissenschaftliche Arbeit, die ich in meinem Studium noch in einem anderen streng wissenschaftlichen Kontext angefertigt habe. Aus diesem Grund nehme ich mir das Recht raus, auf streng wissenschaftliche Stils (jeden Pups vergleichen, strenges Zitieren, super technisches Formulieren etc.) weitesgehend zu verzichten. Natürlich halte ich alle WissenschaftlerInnen und ihre Beiträge an die Wissenschaften in allen Ehren, aber man kann es mit dem Zitieren und Vergleichen auch arg übertreiben. Ich wüßte auch gar nicht wie ich Euklids Arbeiten bezüglich der euklidischen Abstandsnorm zitieren könnte. Gibt es überhaupt eine ISBN die auf das original Papyrus verweist?...

Trotz des Formats, ist diese Arbeit keine Präsentation, sondern tatsächlich eine Zusammenfassung des Projekts um das es eben gehen soll. Ich habe mich aber für dieses Format entschieden, da ich meine, dass es einfach wesentlich effizienter ist als das uralte Hochkantformat. In Zeiten in denen geschätzt 90% der LeserInnen irgendwelcher wohlgeordneter Aneinanderreihungen von Buchstaben Tablets oder Smartphones benutzen, könnte man mal das Hochkantformat langsam etwas überdenken. Der Vorteil ist nämlich der, dass größere Grafiken mit einem ungünstigen Längen-Breitenverhältnis oft viel zu klein im Hochformat platziert werden müssen. Ich meine, dass das "Text- zu Abbildungsmengenverhältnis" in diesem Format so etwas besser ist. Mein kleines Experiment hiermit ist nun - ganz nebenbei - dieser Diktatur des Hochkantformats Einhalt zu gebieten und den weltweit revolutionären Paradigmenwechsel des Breitkantformats einzuleuten

...die LeserInnen werden nun langsam merken, dass diese Arbeit also nicht die typische rohe und kalte wissenschaftliche Arbeit ist. Hier und Da darf mal ein Scherz gemacht werden, oder? Zudem werde ich Dich einfach ganz frech Dutzen, auch wenn wir uns gar nicht kennen.

Wer also dabei sein will und nicht schon jetzt von meinem eigenwilligen unakademischen Schreibstil total angewidert ist, ist herzlich eingeladen weiter zu lesen. Es wird hier und Da auch etwas mathematischer werden... hey, wir machen hier Datenanalyse! Das ist Mathematik in ihrer - aus meiner Sicht - schönsten Form → angewandte Mathematik. Lass Dich aber bitte nicht davon abschrecken. So heftig wird es nicht werden, und es werden fast keine Gleichungen vorkommen ;).

Die gesamte Arbeit ist von A bis Z von mir. Also auch Layout-Design, Satz, die meisten Abbildungen aber auch die mit Sicherheit zahlreichen "orthografische Anomalien". Und auch Interpunktions ist nicht so ganz meine Stärke. Bitte sieh großzügig darüber hinweg. Ein Lektorat konnte ich mir nicht leisten und dafür ist die Arbeit ja auch umsonst.

Bzgl. der Lizenz und Fehlern:

Der Fehlerteufel ist ein fieser Geselle und wird sicherlich auch in dieser Arbeit ordentlich zugeschlagen haben. Oft kann man dem Typen aber nur kollegial gemeinsam entgegen treten. Natürlich habe ich nach bestem Gewissen versucht Fehler zu vermeiden, kann es aber nicht garantieren. Also hilf mir doch bitte gerne, das Ganze noch besser zu machen und teile mir die Fehler freundlich mit.

Du kannst meine Arbeit sehr gerne vervielfältigen und mit anderen Teilen, ohne aber damit Kohle zu machen...ich will schließlich auch keine dafür und hatte nebenbei bemerkt die ganze Arbeit. Meine Abbildungen darfst Du gerne benutzen, aber nur unter Angabe der Quelle und des Autornamen. Unter Einhaltung der bereits genannten Attribute dürfen Inhalte von Dir verändert werden. Für den offiziellen Wortlaut meiner Lizenz klicke bitte einfach das CC-Piktogramm in der rechten unteren Bildecke.

Vielen Dank und viel Spass beim weiteren Lesen,

Alex



Creative Commons  
Namensnennung - Nicht-kommerziell - Weitergabe unter gleichen Bedingungen  
4.0 International Lizenz

# Etwas in Sachen Quellen und Verlinkung

Wie bereits im Vorwort angekündigt, werde ich in dieser Arbeit äußerst sparsam in Sachen Zitieren und Vergleichen sein. Dennoch gibt es hier und da mal einen Verweis. Ein Literaturverzeichnis wirst Du allerding vergeblich suchen. Verweise passieren nämlich als Links im Fließtext. Das macht es Dir wesentlich einfacher zu den Quellen zu gelangen...und ich muss nicht das - bei mir sehr unbeliebte - Literaturverzeichnis anlegen...manchmal bin ich nämlich auch etwas faul.

Es wird Verlinkungen auf Seiten geben, welche angekratzte Sachverhalte noch etwas genauer beleuchten. Dabei wird es sich aber nicht etwa um die „Urquellen“ der Sachverhalte handeln, sondern um - aus meiner Sicht - gut zusammen getragenes Wissen durch andere Autoren. Teils verweise ich auf populärwissenschaftliche Medien. Hier und da gehen die Quellen aber über das Populärwissenschaftliche hinaus. Du kannst sie ja dann schnell wieder weg klicken, wenn's Dir etwas zu abstrakt ist ;).

Populärwissenschaftliche Medien sind, wenn sie gut aufbereitet sind, überhaupt nichts Schlechtes. Und naturwissenschaftliche Sachverhalte durch WIKIPEDIA-Artikel zu bereichern, löst bei manchen Leuten schon regelrechtes Entsetzen aus, aber es ist aus meiner Sicht überhaupt nichts Schlimmes. Vor allem dann nicht, wenn man selber „nur“ eine populärwissenschaftliche Arbeit verfasst. Denn sicherlich gehört sowas nicht in eine rein wissenschaftliche Arbeit. Dort sollten/müssen natürlich die Urquellen des zugrunde liegenden Wissens genannt werden, aber ich möchte alles eben eher populärwissenschaftlich halten und zudem kann man bei Wikipedia auch tiefer gehende Quellen finden.

Du glaubst dennoch nicht, dass WIKIPEDIA in wissenschaftlichen Sachen eine möglicherweise gute bzw. valide Quelle ist? Dann mache doch mal ein Experiment und starte einen eigenen Artikel mit vermeintlich wissenschaftlichen Inhalt und vergesse Quellen und/oder schreibe Unwahrheiten... Es wird garantiert nicht lange auf sich warten lassen, und Dein Artikel fliegt Dir um die Ohren und wird gesperrt oder zumindest markiert werden, weil x Leute mit viel Know-How ihn gelesen und für schlecht befunden haben. Ein härteres Peer-Review-Verfahren als bei WIKIPEDIA gibt es nämlich nicht! Ich wage sogar zu behaupten; es gibt mehr „Bullshit Paper“ - erschienen auf den seriösesten Verlagen - , als unwahre WIKIPEDIA-Artikel.

Hier und da werde ich, wie bereits erwähnt, Links auf Seiten platzieren, welche die Sachverhalte nochmal umfangreicher und - aus meiner Sicht - gut zusammen getragen/aufbereitet haben. Ich kann dennoch nicht für die Richtigkeit dieser Seiten garantieren. Ich übernehme keine Verantwortung/Haftung für Irgendetwas im Zusammenhang dieser Links! Diese Links sind von mir aber nach besten Gewissen ausgesucht wurden und waren zumindest beim Verfassen dieser Arbeit valide. Sie sind als gewohnt blau und unterstrichen „gedruckte“ Wörter im Fließtext deklariert und sofern sie nicht ausgeschrieben sind, noch zusätzlich mit einer Fußnote versehen.

Ein kleines weiters Feature dieser Arbeit sind interne Verlinkungen. So kommst Du etwa vom Inhaltsverzeichnis aus via Klicken interner Links direkt zum Kapitelanfang des gewählten Kapitels. Klicke einfach im Inhaltsverzeichnis auf das Wort des Kapitels und...tada...bist Du auch schon da. Möchtest Du aus einer Seite, die Du gerade liest direkt wieder in das Inhaltsverzeichnis, so klicke einfach auf das „Häuschen“-Symbol in der unteren rechten Seitenecke. Zudem gibt es im Fließtext Verlinkungen innerhalb von verschiedenen Inhalten bzw. derer Seiten. Diese Links sind als grüne unterstrichene Wörter deklariert.

# Motivation des dieser Arbeit Zugrunde liegenden Projekts

Ich bin angehender Data Scientist und ein totaler Zeitreihenanalyse-Freak. Im Studium besuche ich ein Modul, das sich mit Datenvisualisierung und Analysekonzepten beschäftigt. Der Dozent ist ein ziemlich cooler Prof. mit einem klasse didaktischem Konzept. Neben der Vorlesung gibt es eine Übung. Die Übung ist so gestaltet, dass der Dozent uns mit Daten versorgt und wir Studies erstmal alleine irgendwelche Anwendungen nach eigenem Gusto damit machen sollen, wobei natürlich das zuvor in der Vorlesung Gelernte beachtet/angewendet werden soll. In der nächsten Übungsstunde können dann alle Studies ihre Ergebnisse präsentieren und sowohl unser Dozent als auch die Studies geben anschließendes Feedback. Ich finde diesen Ansatz extrem klasse und lehrreich.

Nun gab uns unser Dozent eine Zeitreihe von Wetterdaten über einen Zeitbereich von 1880 bis aktuell. Der Datensatz enthält die üblichen Verdächtigen jener Parameter, welche ein Meteorologenherz höher schlagen lassen. Mittlere Tagestemperatur, Sonnenstunden, Wolkenbedeckungsgrad, Luftdruck etc. Wir sollten uns nun für 3 festlegen und eben ein wenig damit rumexperimentieren. Um ehrlich zu sein war der erste Gedanke, noch bevor ich die Daten gesehen hatte; ich lasse über die Temperaturzeitreihe eine lineare Regression laufen, welche mit Sicherheit die Erderwärmung abbilden wird und fertig ist die Übung...mit einem Aufwands-Koeffizient von geschätzt 0.7 (von 0 bis 10). Aber als ich die Magnituden der mittleren Tagestemperatur über den gesamten Zeitbreich gesehen habe und ihre wunderschönen jährlich-periodischen Verläufe, war die lineare Regression passé. Das ganze hat mich dann so sehr gepackt, dass ich aus der anfänglichen Übungslösung eben ein doch recht umfangreiches Privatprojekt außerhalb des Studiums gemacht habe.

Periodische Prozesse, gleich welcher Art, üben auf mich eine ziemliche Faszination aus. Und nun kann sich jeder denken, dass die hierbei betrachtete Periode eben auf ein Jahr bezieht. Klar, im Sommer ist es warm und im Winter kalt. Und das jedes Jahr. Und das ganze kommt doch bestimmt daher, dass die Sonne im Sommer länger scheint, als im Winter...das lassen wir jetzt erstmal so stehen...stellen uns aber mal bei ruhiger Minute die Frage, warum im Polarsommer/-Winter, wenn nahe den Polen fast 24h die Sonne scheint, dennoch tiefste Temperaturen herrschen ;). Wir werden aber sehen, dass es einen kleinen Zeitversatz zwischen der Sonnenscheindauer und der mittleren Tagestemperatur gibt. Denn um es direkt vorweg zu nehmen, ist der Juni zwar der Monat mit der längsten Sonnenscheindauer -klar der 21. Juni ist schließlich der längste Tag, oder? Aber tatsächlich ist der Juli der wärmste Monat bezogen auf die mittlere Tagestemperatur.

Das Ganze scheint also irgendwie kontraintuitiv zu sein. Und dem wollen wir nun mit ein paar Analysentechniken auf die Spur kommen und beweisen, dass es auch wirklich so ist.

# Motivation dieser Arbeit

Ich möchte Konzepte aus dem Bereich Datenanalyse und Data Science anhand von Anwendungsbeispielen anschaulich und verständlich rüber bringen. Und mit "Vielen" meine ich von SchülerInnen (gut...die 10. Klasse sollte schon besucht wurden sein ;)) bis allen, die sich eben schlicht dafür Interessiert und zumindest etwas Grundverständniss mitbringt.

Ganz ehrlich, die tiefgründige Mathematik hinter all den auftauchenden Konzepten ist enorm kompliziert und ich selber verstehen sie auch nur zu einem geringen Umfang. Aber man muss es auch gar nicht so sehr kompliziert machen.

Wenn man Mathematik auf bestimmten Anwendungsebenen betrachtet, kann sie aus meiner Sicht wirklich für jeden super interessant sein. Aber genau das ist der Punkt. Es geht um ihre Anwendung! Und genau das wird - aus meiner Erfahrung und Sicht – in der Schule unzureichend vermittelt. Oder hat Dir Mathe Spaß gemacht und findest sie interessant?

Ich hatte damals in der 10. Klasse auf dem Zeugnis eine 5!!! in Mathe zu stehen... und nun studiere ich quasi angewandte Mathematik! Hätte mir meine Lehrerin damals gezeigt, dass man lineare Funktionen als klasse Modellrepräsentationen zum Beschreiben irgendwelcher interessanter Prozesse (Weg-Zeit-Gesetz bei konstanter Geschwindigkeit, lineares Wachstum etc.) nutzen kann, als einfach nur unbedeutende Linien stumpf in irgendwelche Koordinatensysteme zeichnen und Anstiege anhand von zwei absolut nichtssagenden Punkten zu berechnen zu lassen, hätte das vielleicht etwas anders ausgesehen.

Mathematik ist nicht unbedingt etwas total Abstraktes. Klar in ihrer Reinform ist sie das. Und das ist auch gut so, denn das muss sie sogar! Denn das macht sie zu diesem universellen und mächtigen Werkzeug. Aber man kann sie eben auch allgemeiner und vor allem anschaulicher betrachten, und ihr den Charm entlocken der eben auch in ihr steckt.

Datenanalyse ist - aus meiner Sicht - eine der schönsten und anschaulichsten Formen angewandter Mathematik. Wir können interessante Informationen aus Daten gewinnen. Wenn wir diese Informationen auch noch geschickt und verständlich visualisieren und interpretieren, kann sich so ziemlich jeder daran faszinieren.

## ***Mein Appell an die MathelehrerInnen:***

In Zeiten vorranschreitender Digitalisierung brauchen wir zunehmend InformatikerInnen and Data Scientists. Vergrault eure SchülerInnen nicht mit wenig anschaulichen Beispielen. Vergält es ihnen nicht schon vor einem Studium mit übertriebenen Abstraktionen, sondern weckt, durch interessante Beispiele, Veranschaulichungen und Experimente das potenzielle Interesse der SchülerInnen!

# Inhalt

Einleitung

Datengrundlage, Sichtung und Hypothese

Methodik

Preprocessing

Kleiner Exkurs in die Welt des (gleitenden) Mittelwerts

Methode I – Gleitender Mittelwert und Maximum-Detektion

Methode II – Anpassung einer Sinusfunktion

Methode III – Lissajous-Ansatz

Alle Modelle und Ergebnisse im Vergleich

Hypothesentest

Lessons Learned

Ergebnisinterpretation

# Einleitung

Das Wetter unterliegt bekanntermaßen einem jährlichen Rhythmus. Dabei schwanken die Temperaturen periodisch zwischen Kalt (Winter) und warm (Sommer) sowie aber auch die Sonnenscheindauer periodisch schwankt. Im Winter haben wir kürzere Tage als im Sommer. Ich beziehe mich dabei auf die Nordhalbkugel der Erde.

Dies lässt uns annehmen, dass die Temperatur direkt mit der Sonnenscheindauer zusammenhängt. Wenn es eine (gleichmäßige) Schwankung bei beiden Parametern gibt, dann müsste nach dieser Logik für jeden Parameter ein zeitliches Maximum geben, welches für beide Parameter dann ja auch noch auf einen gleichen Zeitpunkt fallen sollten. Also wenn Tag X immer der längste ist, dann sollte er doch auch immer der wärmste sein...oder? Nun, zwar verhalten sich beide Parameter periodisch und sie sind zweifelsohne miteinander verknüpft, aber die vermeindliche Logik, dass der Tag mit der längsten Sonnenscheindauer auch der wärmste ist, ist nicht richtig. Es gibt einen zeitlichen Versatz zwischen beiden Parametern.

Mithilfe eines umfangreichen Datensatzes und einiger Analyse-Methoden werden wir dem Ganzen im Folgenden auf die Spur kommen. Bei dieser Gelegenheit testen wir verschiedene Filtermethoden auf ihre Anwendbarkeit und werten sie statistisch aus. Wir benutzen Ansätze aus dem Bereich des maschinellen Lernens, also künstliche Intelligenz wie dem DBSCAN-Clustering-Algorithmus und Regressionsanalysen und trainieren Modelle auf Daten, um mit ihnen dann Erkenntnisse zu gewinnen bzw. unsere Hypothese des bereits erwähnten zeitlichen Versatzes beider Wetterparameter zu beweisen. Das Ganze werden wir zudem auch noch statistisch sinnvoll validieren, um uns unserer Hypothese bzw. vielmehr ihrer Gültigkeit auch wirklich sicherer sein zu können

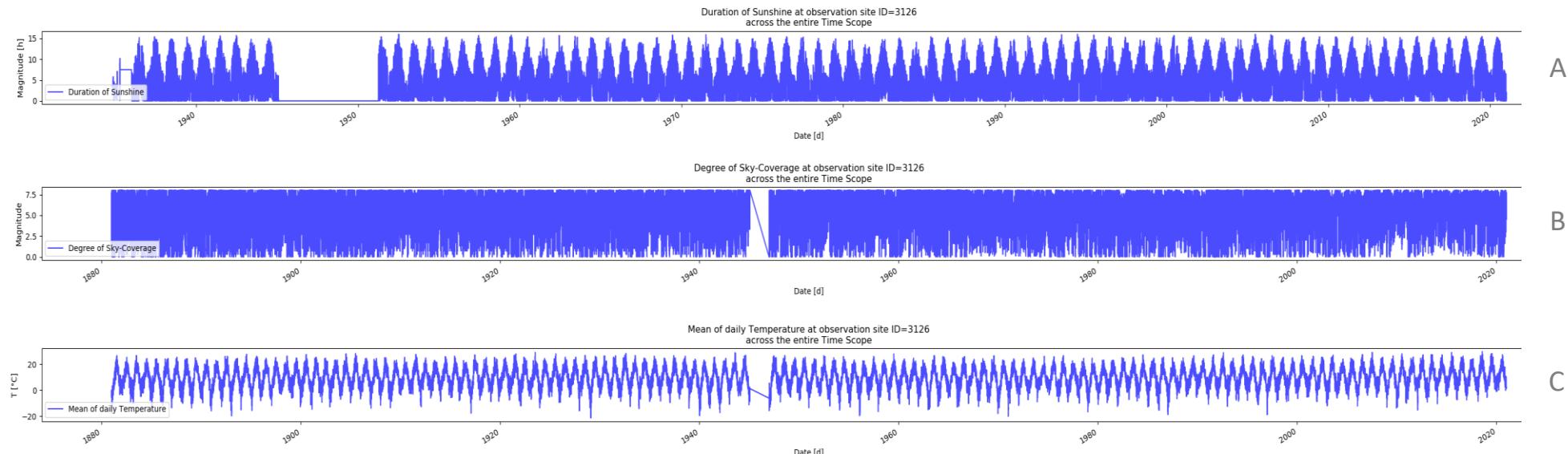
Wir werden uns dazu einige interessante Filtertechniken der Signalverarbeitung anschauen und dabei bei der letzten Methode sogar eine, zugegeben, sehr exotische Methode in unserm Kontext prüfen. Diese ist das i-Tüpfelchen des gesamten Projekts und wenn auch nicht ganz so performant - wie wir sehen werden - zumindest aber recht kreativ und funktionierend...wenn ich mir diesen gewaltigen Eigenlob mal eben erlauben darf.



# Datengrundlage, Sichtung und Hypothese

Als Datengrundlage dient uns ein Wetterdatensatz welcher verschiedenste Wetterparameter zeitlich auf den Messstandort Magdeburg abbildet. Das Messintervall beträgt für jeden Parameter einen Tag. D.h. für jeden Tag existiert - im Idealfall - ein zugehöriger Messwert.

Auch wenn der eigentliche Datensatz sehr viele weitere Parameter enthält, konzentrieren wir uns aber im gesamten Projekt nur auf zwei Parameter. Die folgende Abbildung (Abb.1) ist eine Datenvisualisierung der Messparameter Sonnenscheindauer (A), des Wolkenbedeckungsgrades (B) und der mittleren Tagestemperatur (C). Das Ganze geht teilweise über einen Zeitraum von 1880 bis 2020. Zu sehen sind einige sehr große Lücken, etwa bei A zwischen ca. 1945 bis 1947. Auch bei B und C sieht man eine größere Lücke in diesem Bereich. Man kann jetzt wild herum spekuliere, den offensichtlich spielt es sich das Ganze in einem Zeitbereich kurz nach zweiten Weltkriegs ab, aber es spielt eigentlich auch keine weitere Rolle. Denn unser Vorgehen wird sein, die Daten ab 1955 bis Ende zu nehmen. Dann brauchen uns diese Lücken nicht mehr zu interessieren und wir können uns zudem etwas sicherer sein, dass modernere Messgeräte für die Messungen genutzt wurden. Auf den ersten Blick sieht man auf jeden Fall schon sehr gut das periodische Verhalten bei Sonnenscheindauer und Temperatur. Bei dem Bedeckungsgrad dagegen kann man das sicherlich nicht.



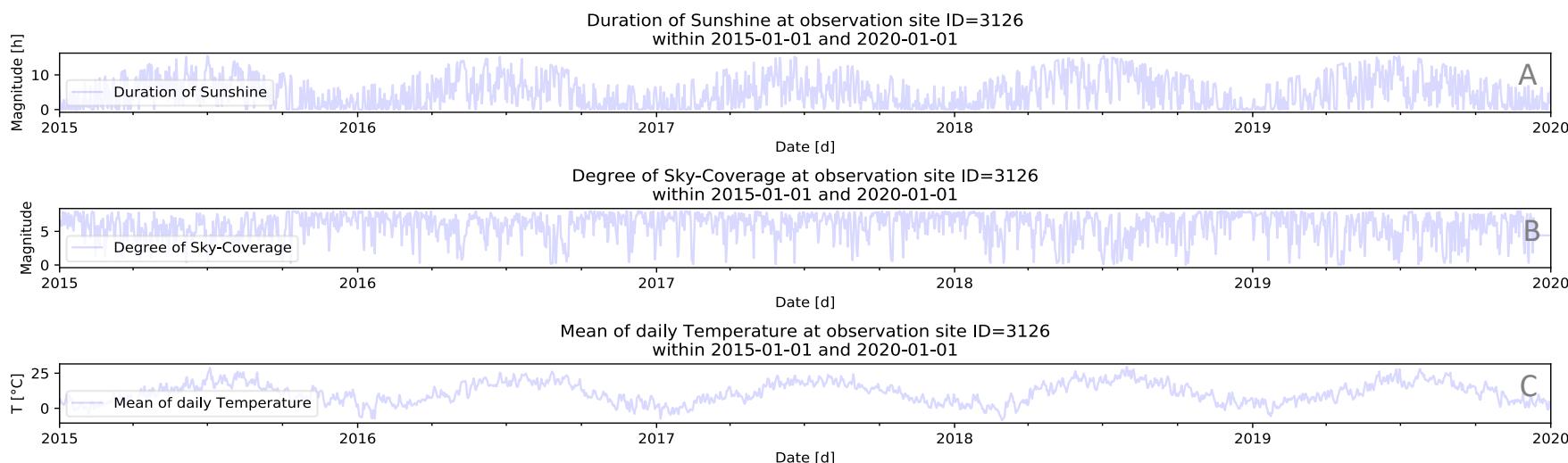
**Abbildung 1:** Grobvisualisierung der Zeitreihendaten Sonnenscheindauer (A), Wolkenbedeckungsgrad (B) und der mittleren Tagestemperatur (C). Bei A und C ist schon in diesem Zeitbereich ein periodisches Verhalten der Magnituden gut zu sehen. Bei dem Bedeckungsgrad (B) kann man dagegen kein periodisches Verhalten erkennen.



# Datengrundlage, Sichtung und Hypothese

Die Abbildung 1 ist zugegeben noch etwas sehr grob gewesen und dient eben auch nur einem ersten groben Überblick. Es kann gesehen werden, dass es Datenlücken gibt und diese entsprechend behandelt werden müssen. Unsere Strategie ist nun die; jüngere Zeitbereiche zwischen 1955 bis 2020 zu nutzen, da in diesem Bereich keine Lücken auf zu treten scheinen, und die Messtechnik vermutlich moderner und genauer ist... (eine zugegeben sehr subjektive Annahme). Der erste Eindruck des periodischen Verhaltens sollte nun noch besser durch eine geeigneter Visualisierung geschehen. Abbildung 2 geht nun etwas detaillierter ran und wir beschränken uns auf einen kleineren Zeitbereich zwischen 2015 und 2020. Ein weiterer Vorteil ist nun zudem der, dass alle Zeitachsen auf diesen Zeitbereich normalisiert sind. Sie liegen also bzgl. ihrer Zeitpunkte exakt übereinander. Das macht uns das Erkennen von möglichen zeitlichen Zusammenhängen der Messparameter schonmal etwas einfacher.

Das periodische Verhalten ist bzgl. A und C noch besser zu erkennen. Ganz vorsichtig darf man sogar auch ein periodisches Verhalten beim Bedeckungsgrad (B) annehmen. Aber wir werden ihn sowieso bald ignorieren. Allerdings können wir ja zumindest die Vermutung anstellen, dass er ebenso jährlich periodisch ist, wie die Temperatur und der Sonnenscheindauer. Denn gerade letztere hängt ja mit ihm ab... wenn es bedeckter ist, dann ist die effektive Sonnenscheindauer eben kürzer.



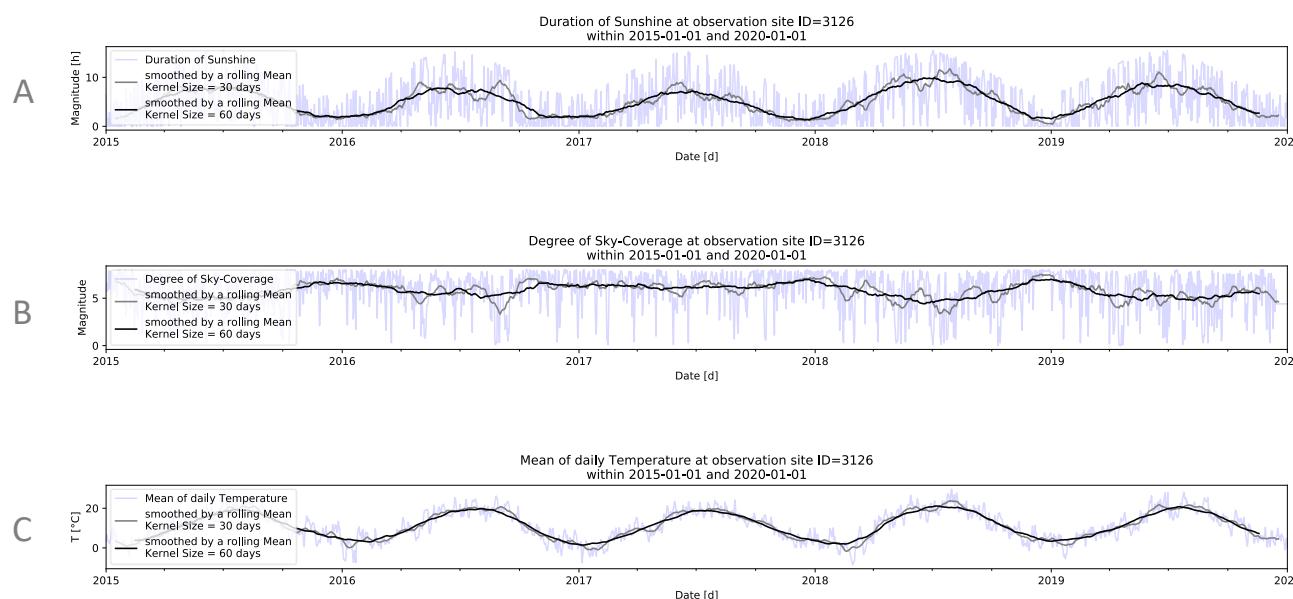
**Abbildung 2:** Genauere Visualisierung der Zeitreihendaten Sonnenscheindauer (A), Wolkenbedeckungsgrad (B) und der mittleren Tagestemperatur (C). Das periodische Verhalten der Magnituden ist bei A und C noch wesentlich besser zu erkennen. Bei B ist ein periodisches Verhalten nun ebenfalls etwas erkennbar. Dabei scheinen A und B antagonistisch zueinander zu sein. Dies sollte jedoch nicht verwundern, da A als effektive Sonnenscheindauer auch vom Bedeckungsgrad (B) abhängt.



# Datengrundlage, Sichtung und Hypothese

Zwar geht es in diesem Abschnitt noch nicht konkret um Datenfilterung als Methode aber wir werden nun dennoch schon einen ersten Filter auf die Daten benutzen um einen möglicherweise noch besseren Eindruck von den Daten und ihrer Beschaffenheit zu bekommen. Die Rede ist von einem Glättungsfilter auf Basis eines gleitenden Mittelwerts. Dieser kann Daten mit einer gewissen Varianz hinsichtlich dieser Varianz glätten. Varianz bedeutet in diesem Fall, dass die Daten etwas variieren. Sie streuen um einen spezifischen Wert. Aber schauen wir uns das einfach mal an. Abbildung 3 ist abgesehen von der zusätzlichen Filterergebnisvisualisierung ansonsten identisch zu Abbildung 2.

Es wurden für jeden Parameter zwei gleitende Mittelwerte mit jeweils unterschiedlichen Fensterlängen angewendet (schwarz: Länge = 60 Tage; grau: Länge = 30 Tage). Insbesondere die schwarze Kurve zeigt für A und C sehr gute Glättungseigenschaften, welche durch eine sehr starke Minimierung der Streuung der Magnitudenwerte zu erkennen sind.



**Abbildung 3:** Genaue Visualisierung der Zeitreihendaten Sonnenscheindauer (A), Wolkenbedeckungsgrad (B) und der mittleren Tagstemperatur (C). In dieser Visualisierung ist nun zusätzlich eine Datenglättung mit einbezogen wurden. Zu sehen sind nun zusätzlich eine schwarze und eine graue Kurve, welche beide die geglätteten Daten repräsentieren. Der Unterschied liegt in der sogenannten Fenstergröße des Glättungsfilters. Es wurden zwei verschiedene Werte genutzt. Zum einen eine Fensterlänge von 30 Tagen (graue Kurve) und 60 Tage (schwarze Kurve).

Denn die graue und vor allem die schwarze Kurve ist in jedem Fall wesentlich glatter.

Bei B wird die angenommene periodische Magnitude in den Jahren 2018-2019 und 2019-2020 ebenfalls nochmal etwa akzentuiert. Ein solcher Glättungsfilter kann also schon bei der Datensichtung ein sehr nützliches Werkzeug sein.

Wir werden uns ohnehin nochmal intensiv mit diesem Filter im Methodenteil befassen.

Nun ist es Zeit eine kleine Hypothese in den Raum zu werfen. Denn nun, nach der Glättung, kann man zumindest schon etwas erahnen, nämlich, dass die Magnituden-Maxima bei A und C nicht exakt übereinander liegen, oder?

Du erkennst es noch nicht?

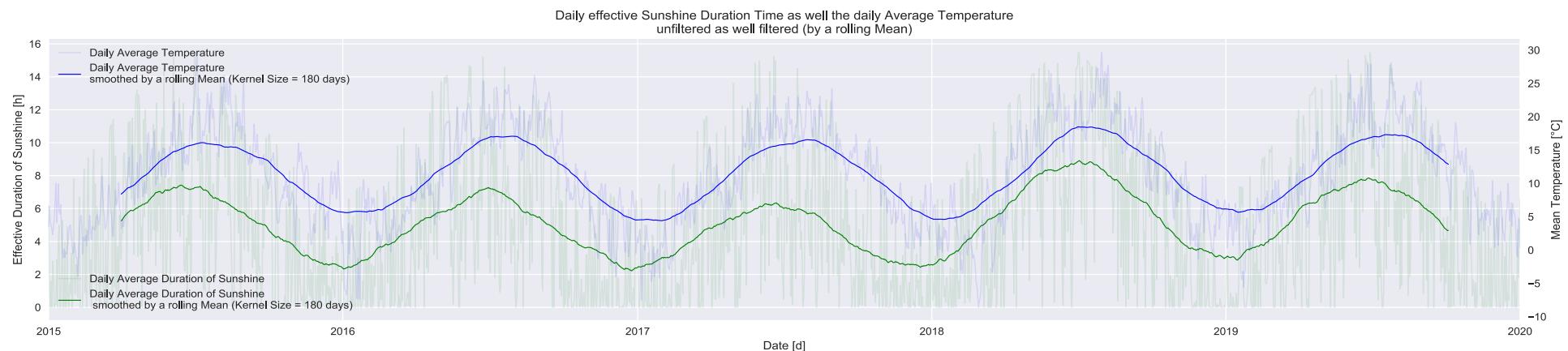


# Datengrundlage, Sichtung und Hypothese

Nachdem die Daten geglättet wurden, kann man wunderbar das periodischen Verhalten hinsichtlich der mittleren Tagestemperatur und der Sonnenscheindauer sehen. Für den Bedeckungsgrad kann man es nun zumindest etwas besser sehen, dass es ein paar zeitliche Passagen innerhalb des visualisierten Zeitbereichs gibt, in denen ein solches periodisches Verhalten bestehen könnte. Ich bin mir ziemlich sicher, dass wir eine Periodizität für diesen Parameter nachweisen könnten, aber wir haben nun schon eine Hypothese und belassen es auch nur bei dieser. Wir konzentrieren uns nun also nur noch auf die beiden Parameter mittlere Tagestemperatur und der Sonnenscheindauer und versuchen nun nachzuweisen, dass sich die Maxima verschoben zueinander verhalten.

Schauen wir uns dazu am besten noch eine etwas detailliertere Abbildung (Abb. 4) an, welche nun nur noch die beiden relevanten Parameter enthält und diese in nur einer Grafik darstellt. Die blaue Kurve repräsentiert die mittlere Tagestemperatur und die grüne die Sonnenscheindauer. Wir benutzen die bereits für gut befundene Glättung mittels des gleitenden Mittelwerts und können deutlich sehen, dass es eine Verschiebung zwischen den Magnituden zu geben scheint. Die Maxima fallen auf unterschiedliche Tage. Und vor allem liegen die Maxima der mittleren Tagestemperatur in allen Fällen rechtsverschoben relativ zu Sonnenscheindauer.

Die Hypothese, dass die Magnituden-Maxima zeitlich zueinander versetzt sind, scheint sich damit schonmal etwas zu erhärten. Nun sollten wir die Hypothese auch vernünftig beweisen. Daten dafür haben wir ja schließlich genug.



**Abbildung 4:** Überlagerungs-Plot der beiden Parameter mittlere Tagestemperatur (blaue Kurve) und tägliche Sonnenscheindauer (grüne Kurve). Anhand dieser Abbildung wird die Hypothese, dass sich die beiden Magnituden-Maxima zeitlich versetzt zueinander verhalten nochmals etwas bestärkt. Beide Wetterparameter liegen sowohl in gefilterter (Glättung) als auch ungefilterter Variante vor. Erst durch die Datenglättung via dem Glättungsfilter ist die Verschiebung der Magnitudenmaxima beider Wetterparameter wirklich gut erkennbar. Für den Filter wurde der gleitende, zentrierte Mittelwert mit einer Fensterlänge von 180 Tagen gewählt. Aus diesem Grund fehlen sowohl am Ende als auch am Anfang des gesamt betrachteten Zeitraums von 2015 bis 2020 jeweils 90 Tage bei den Kurven, welche die geglätteten Daten repräsentieren.



# Methodik

Die Sichtung der Daten mithilfe des Glättungsfilters hat auf einen kleinen Zeitbereich die Gültigkeit der Hypothese stark bekräftigt. Aber statistisch einwandfrei bzw. ein Beweis für ihre Gültigkeit ist das noch lange nicht!

Es kann nämlich sehr gut auch nur reiner Zufall sein, dass sich die Magnitudenmaxima innerhalb des betrachteten Zeitraums verschoben verhalten. Schließlich streuen alle Messungen meistens irgendwie, und wir hatten ja in den ungefilterten Daten gesehen, dass die Magnituden sehr stark streuen.

Da wir aber glücklicherweise einen weitaus größeren Datensatz haben, können wir auch noch weitere Jahre anschauen. Anschauen ist dabei auch schön und gut, aber auch das ist aus statistischer Sicht noch nicht wirklich toll. Denn selbst wenn wir jetzt den gesamten Zeitbereich anschauen, hat das keinen wirklichen statistischen Wert. Und „Wert“ ist genau der Punkt. Wir wollen eben einen statistischen Wert haben, eine Statistik, eine Metrik, etwas, mit dem wir numerisch unsere Hypothese bestätigen (oder eben auch widerlegen) können. Etwa durch Vergleichen mit Referenzwerten.

Uns interessiert also nun nicht nur, ob es viele Jahre gab in denen dieser Effekt auftrat, sondern wie stark und ob es ein klares Muster gibt oder alles eher zufällig passiert. Wenn die Magnituden-Maxima wild gegeneinander verschoben sind, und es dabei kein statistisches Muster gibt, dann müssen wir uns damit abfinden. Noch wissen wir es aber nicht.

Wir erweitern jetzt die Hypothese etwas, nämlich darum, dass die Verschiebung der Magnituden-Maxima sich auf einen bestimmten Zeitpunkt im Jahr beziehen. D.h. wir nehmen an, es gibt für jeden Wetterparameter einen signifikanten Zeitpunkt des Maximums, um den aber die Maxima-bezogenen Tage etwas schwanken dürfen. Die Verschiebung beließe sich dann also auf diese beiden angenommen signifikanten Zeitpunkte in Form einer (signifikanten) Distanz zueinander.

Unsere Aufgabe ist nun eine Analyse vor zu nehmen, um die erwarteten signifikanten Zeitpunkte der Magnituden-Maxima nachzuweisen bzw. zu bestimmen und statistisch auszuwerten. Wir haben 65 Jahre zur Verfügung und sehr viele Datenpunkte innerhalb dieser 65 Jahre. Somit steht einer vernünftige statistischen Auswertung eigentlich schonmal nichts entgegen.

Unser Ziel wird es also sein, Methoden zum Bestimmen der Maximalwerte zu entwickeln um für jeden Wetterparameter seinen signifikanten Zeitpunkt seines Maximums (statistisch valide) zu bestimmen.

Die statistische Auswertung wird dann über ein Monte-Carlo-Kreuz-Validierungs-Verfahren passieren. In den Methodenteilen wird das ganze etwas ausführlicher diskutiert.



# Preprocessing

In Abbildung 1 konnten wir sehen, dass es Zeitbereiche mit sehr großen Datenlücken gibt (siehe dort die Bereiche um 1943). Solche Datenlücken sind nicht in jeden Fall behebbar. Es kommt dabei sehr auf den Kontext an. Aber da wir ohnehin einen jüngeren Zeitbereich betrachten, soll dieses Problem gar keins für uns sein.

Dennoch gibt es auch im betrachteten Zeitbereich kleine Lücken. Oft nicht größer als ein oder zwei Messintervalle. Man kann sie in den Visualisierungen gar nicht erkennen, aber sie existieren. Und was auch immer der Grund für solche Lücken ist (unmotivierte Messingenieure, Messgeräteaussetzer etc.), statistisch gesehen fallen sie gar nicht wirklich ins Gewicht. Dennoch müssen wir sie behandeln, da viele Algorithmen bzw. ihre Implementierung in Programme, die wir zur Analyse nutzen, nicht mit fehlenden Daten umgehen können. Blöd ausgedrückt; der Computer erwartet immer einen Wert, und den müssen wir ihn irgendwie geben. Wir müssen also Ersatzwerte finden. Es gibt Algorithmen die mit NULL-Werten umgehen können. Das sind quasi gar keine Werte, jeden Falls nicht aus numerischer Sicht, aber sie funktionieren manchmal als Ersatzwert. In den meisten Fällen braucht es aber „echte“ Werte.

Um nun sinnvolle Ersatzwerte zu finden, kann man sich bestimmter Methoden zum Auffüllen solcher Datenlücken bedienen. Man kann z.B. den letzten bekannten Wert, also den Nachbarwert benutzen und den fehlenden Wert durch diesen ersetzen. Oder man nutzt zum Ersetzen den Mittelwert über alle Werte aus dem Datensatz. Ich persönlich präferiere den Nachbarwert bei periodisch stationären Signalen auf kleinen Zeitbereichen. Es erscheint mir einfach wahrscheinlicher, dass der Nachbarwert näher am fehlenden Wert liegt, als das es der Mittelwert aller Werte würde. Wir könnten diese kleine Behauptung, dass es zumindest bei periodischen stationären Signalen wirklich so ist, sogar statistisch beweisen. Und im streng wissenschaftlichen Kontext, wo jeder Methodenteil streng validiert werden muss/sollte, müsste ich es sogar tun, aber vertrau mir jetzt einfach, dass es so ist.

Und im Übrigen ist das Verhältnis zwischen fehlenden Werten zur Gesamtanzahl der Zeitpunkte über den ganzen Beobachtungszeitraum so klein, dass der Einfluss auf die Analyseergebnisse wirklich sehr gering sein dürfte, egal welche Methode man zum Auffüllen denn nun nehmen würde.

Weitere Vorprozessierungsschritte werden in den einzelnen Methodenteilen erklärt, da sie teils quasi selber Teil der Methode ansich sind.



# Kleiner Exkurs in die Welt des (gleitenden) Mittelwerts

...und warum ein uneindeutiger Umgang mit dem Begriff Mittelwert zu Verwirrungen und/oder falschen Eindrücken führen kann.

Zunächst muss ich erstmal um Entschuldigung dafür bitten, dass ich mit dem Begriff Mittelwert – ob nun gleitend oder nicht – sehr frevelhaft umgehe und auch weiterhin umgehen werde.

Denn in der Statistik gibt es so einige verschiedene Mittelwerte. Der Begriff ist nämlich alles andere als Eindeutig. Und dass das nicht nur für Verwirrungen, sondern auch zu falsche Eindrücken führen kann, werden wir noch in einem kleinen Beispiel hierzu sehen. Wenn ich nun aber den (gleitenden) Mittelwert oder aber auch Durchschnitt als Begriff benutze, dann meine ich damit generell immer den arithmetischen Mittelwert, wenn ich den Median als eine ähnliche Metrik verwende, nenne ich ihn auch so!

Salopp gesagt ist der Mittelwert der, den wir alle aus der Schule und Alltag kennen. Wir haben einen Topf voller Werte und addieren diese Werte einfach auf und teilen diese Summe noch durch die Anzahl dieser Werte. Aber daneben gibt es z.B. noch den Median. Das ist der Wert, der sich in der „Mitte“ befindet, wenn Du alle Werte aus dem Topf der Reihe nach, nach ihrer Größe sortierst. Nebenbei gibt es aber auch noch den gewichteten Mittelwert, harmonischer Mittelwert, geometrischer Mittelwert ...etc. Aber auf die möchte ich gar nicht weiter eingehen.

Was ist nun der gleitende Mittelwert?

Unsere Zeitreihen, etwa die der Temperaturen, besteht aus vielen Messwerten. Wir könnten nun aus allen Messwerten den Mittelwert nach dem bereits bekannten Prinzip berechnen. Wenn wir diesen Wert jetzt aber auf die Zeit abbilden würden, dann gibt es zu jedem Zeitpunkt immer nur diesen einen Wert. Wir hätten einfach eine horizontale Linie im Diagramm, welche nebenbei bemerkt wenig Aussagekraft für Irgendwas hätte. Diese tolle Datenglättung, welche ja dennoch unterschiedliche Werte auf die korrespondierenden Zeitpunkte abbildet, bekommt man damit also schonmal nicht hin. Die Idee dahinter ist nämlich eine etwas Andere. Man betrachtet immer nur einen kleinen zeitlichen Ausschnitt (ein Fenster) aus der gesamten Datenreihe um einen korrespondierenden Zeitpunkt und bildet aus den darin befindlichen Werten den Mittelwert. Dann lässt man das Fenster um einen Zeitschritt weiter wandern (gleiten) und wiederholt die Prozedur. Zu jedem korrespondierenden Zeitschritt gibt es also immer einen eigenen Mittelwert, welcher aus dem gleitenden Fenster um diesen Zeitpunkt errechnet wurde – bzw. vielmehr aus den Werten innerhalb dieses Fensters. Es gibt dabei noch die Unterscheidung ob das Fenster am Zeitschritt liegt und dabei zurück oder voraus „blickt“ oder eben um den Zeitschritt (zentrierter gleitender Mittelwert). Aber so tief wollen wir uns damit nun auch nicht beschäftigen. Auf den nächsten Seiten soll das Prinzip des gleitenden Mittelwerts etwas bildhafter erläutert werden.

Die obere Leiste mit den ganzen  $a$  repräsentiert etwas abstrakt unsere Zeitreihe.  $a$  ist der Parameterwert und die tiefgestellte Ziffer der korrespondierende Zeitpunkt zum Parameterwert. Das Fenster hat in diesem Beispiel die Länge 5 und wandert nun pro Buchseite um einen Zeitschritt weiter und berechnet jeweils den „Glättungswert“  $b$  zu dessen korrespondierenden Zeitschritt. Es wird durch den roten gleitenden Balken dargestellt. Dir wird auffallen, dass die ersten beiden Glättungswerte ( $\backslash_1$  und  $\backslash_2$ ) nicht wirklich Werte sind. Sie existieren beim zentrierten gleitenden Mittelwert nicht. Genauso wie zu den beiden letzten Zeitpunkten keine berechnet werden können. Sie lassen sich schließlich nicht berechnen. Aber Du wirst es gleich selber verstehen.



# Kleiner Exkurs in die Welt des (gleitenden) Mittelwerts

...und warum ein uneindeutiger Umgang mit dem Begriff Mittelwert zu Verwirrungen und/oder falschen Eindrücken führen kann.

$a_1$	$a_2$	$a_3$	$a_4$	$a_5$	$a_6$	$a_7$	$a_8$	$a_9$	$a_{10}$	$a_{11}$	$a_{12}$	$a_{13}$	$a_{14}$	$a_{15}$	$a_{16}$	$a_{17}$	$a_{18}$	$a_{19}$	$a_{20}$	$a_{21}$	$a_{22}$	$a_{23}$	$a_{24}$	$a_{25}$	$a_{26}$	$a_{27}$	$a_{28}$	$a_{29}$	$a_{30}$	$a_{31}$	$a_{32}$	$a_{33}$	$a_{34}$	$a_{35}$	$a_{37}$	$a_{38}$	$a_{39}$	$a_{40}$	$a_{41}$
-------	-------	-------	-------	-------	-------	-------	-------	-------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------

$$\frac{a_1 + a_2 + a_3 + a_4 + a_5}{5}$$

$\backslash_1 \quad \backslash_2 \quad b_3$

[Klick mich](#)

# Kleiner Exkurs in die Welt des (gleitenden) Mittelwerts

...und warum ein uneindeutiger Umgang mit dem Begriff Mittelwert zu Verwirrungen und/oder falschen Eindrücken führen kann.

$a_1$	$a_2$	$a_3$	$a_4$	$a_5$	$a_6$	$a_7$	$a_8$	$a_9$	$a_{10}$	$a_{11}$	$a_{12}$	$a_{13}$	$a_{14}$	$a_{15}$	$a_{16}$	$a_{17}$	$a_{18}$	$a_{19}$	$a_{20}$	$a_{21}$	$a_{22}$	$a_{23}$	$a_{24}$	$a_{25}$	$a_{26}$	$a_{27}$	$a_{28}$	$a_{29}$	$a_{30}$	$a_{31}$	$a_{32}$	$a_{33}$	$a_{34}$	$a_{35}$	$a_{37}$	$a_{38}$	$a_{39}$	$a_{40}$	$a_{41}$
-------	-------	-------	-------	-------	-------	-------	-------	-------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------

$$\frac{a_2 + a_3 + a_4 + a_5 + a_6}{5}$$



$\backslash_1 \quad \backslash_2 \quad b_3 \quad b_4$

[Klick mich](#)

# Kleiner Exkurs in die Welt des (gleitenden) Mittelwerts

...und warum ein uneindeutiger Umgang mit dem Begriff Mittelwert zu Verwirrungen und/oder falschen Eindrücken führen kann.

$a_1$	$a_2$	$a_3$	$a_4$	$a_5$	$a_6$	$a_7$	$a_8$	$a_9$	$a_{10}$	$a_{11}$	$a_{12}$	$a_{13}$	$a_{14}$	$a_{15}$	$a_{16}$	$a_{17}$	$a_{18}$	$a_{19}$	$a_{20}$	$a_{21}$	$a_{22}$	$a_{23}$	$a_{24}$	$a_{25}$	$a_{26}$	$a_{27}$	$a_{28}$	$a_{29}$	$a_{30}$	$a_{31}$	$a_{32}$	$a_{33}$	$a_{34}$	$a_{35}$	$a_{37}$	$a_{38}$	$a_{39}$	$a_{40}$	$a_{41}$
-------	-------	-------	-------	-------	-------	-------	-------	-------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------

$$\frac{a_3 + a_4 + a_5 + a_6 + a_7}{5}$$



$\setminus_1$	$\setminus_2$	$b_3$	$b_4$	$b_5$
---------------	---------------	-------	-------	-------

[Klick mich](#)

# Kleiner Exkurs in die Welt des (gleitenden) Mittelwerts

...und warum ein uneindeutiger Umgang mit dem Begriff Mittelwert zu Verwirrungen und/oder falschen Eindrücken führen kann.

$a_1$	$a_2$	$a_3$	$a_4$	$a_5$	$a_6$	$a_7$	$a_8$	$a_9$	$a_{10}$	$a_{11}$	$a_{12}$	$a_{13}$	$a_{14}$	$a_{15}$	$a_{16}$	$a_{17}$	$a_{18}$	$a_{19}$	$a_{20}$	$a_{21}$	$a_{22}$	$a_{23}$	$a_{24}$	$a_{25}$	$a_{26}$	$a_{27}$	$a_{28}$	$a_{29}$	$a_{30}$	$a_{31}$	$a_{32}$	$a_{33}$	$a_{34}$	$a_{35}$	$a_{37}$	$a_{38}$	$a_{39}$	$a_{40}$	$a_{41}$
-------	-------	-------	-------	-------	-------	-------	-------	-------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------

$$\frac{a_4 + a_5 + a_6 + a_7 + a_8}{5}$$



$\setminus_1$	$\setminus_2$	$b_3$	$b_4$	$b_5$	$b_6$
---------------	---------------	-------	-------	-------	-------

[Klick mich](#)

# Kleiner Exkurs in die Welt des (gleitenden) Mittelwerts

...und warum ein uneindeutiger Umgang mit dem Begriff Mittelwert zu Verwirrungen und/oder falschen Eindrücken führen kann.

und so weiter...



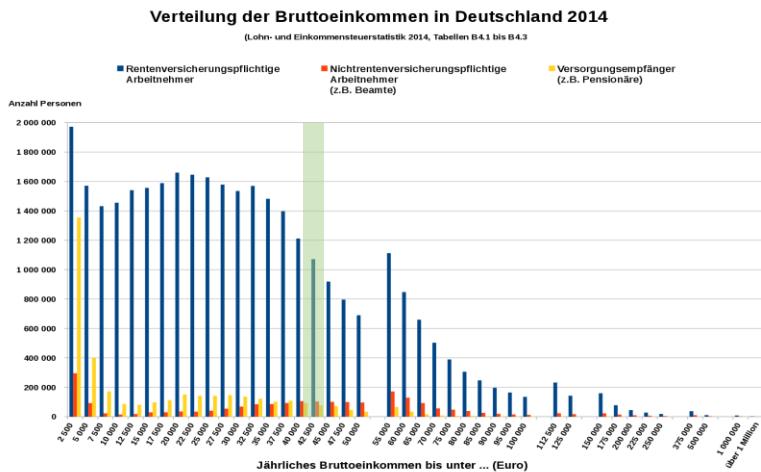
# Kleiner Exkurs in die Welt des (gleitenden) Mittelwerts

...und warum ein uneindeutiger Umgang mit dem Begriff Mittelwert zu Verwirrungen und/oder falschen Eindrücken führen kann.

Und jetzt noch ganz schnell was zum Thema Mittelwert und der Problematik bzgl. der Begrifflichkeit.

Leider finde ich nichts dazu im Netzt, aber ich meine mal gelesen oder gehört zu haben, dass eine sehr bekannte Politikerin mal davon sprach, dass es den Bürgern der BRD gar nicht so schlecht ginge, da das Durchschnittseinkommen schließlich über 3500 €/Monat liegen würde... die Rede ist hier von einer Politikerin die zudem noch promovierte Quantenphysikerin ist und daher wirklich sehr viel Ahnung von Statistik haben wird. Aus meiner Sicht war das schon deshalb eine sehr gewagte Aussage.

Denn zwar ist der Teil der Aussage bzgl. des Betrags des Durchschnittseinkommens sicherlich richtig, aber ein arithmetisches Mittelwert – und dieser wird häufig bei dieser Einkommensstatistik genannt – verzerrt leider das Ganze, wie wir gleich sehen werden, etwas ...arg. Ein arithmetischer Mittelwert funktioniert anschaulich gesehen nämlich nur gut bei normal- oder ähnlich geartet verteilten Daten. Also bei Daten die „symmetrisch“ um ihren arithmetischen Mittelwert verteilt liegen. Wir sehen aber in der Abbildung unten (Zusatzabbildung i), dass die Verteilung der Einkommen in Deutschland eine unsymmetrische Verteilung ist.



**Zusatzabbildung i:** Histogramm über das Bruttoeinkommen verschiedenartig versicherter Arbeitnehmer in Deutschland im Jahr 2014. Verändert durch den Autor.  
Originalbildbeschreibung: Die Grafik zeigt die Verteilung der Bruttoeinkommen der Arbeitnehmer in Deutschland 2014.  
Originaldatenquelle: Statistisches Bundesamt, Fachserie 14 Reihe 7.1, Lohn- und Einkommensteuer, Tabellen B4.1, B4.2 und B4; [UDO BRECHTEL, 2019]; CC-BY-SA  
Entnommen aus Wikipedia. Quelle:  
[https://commons.wikimedia.org/wiki/File:Verteilung\\_Bruttoeinkommen\\_2014\\_in\\_Deutschland.svg](https://commons.wikimedia.org/wiki/File:Verteilung_Bruttoeinkommen_2014_in_Deutschland.svg)

Die Abbildung zeigt ein Histogramm, welches die Häufigkeit eines bestimmten Einkommensbereichs, einer Klasse, auf einen Balken bzgl. seiner Länge abbildet. Die Grafik unterscheidet anhand der Balkenfarben die einzelnen Versicherungsarten. Uns interessiert besonders die Rentenversicherungspflichtigen Arbeitnehmerstatistik (blaue Balken). Das ganze passiert nun in 2500 €/Jahr Abstufungen.

Der erste Balken entspricht also 2500/Jahr und ist der Längste, da in dieser Einkommensklasse die meisten Sozialversicherungspflichtigen ArbeitnehmerInnen zu finden sind. Die Häufigkeit kannst Du auf der Y-Achse ablesen. Der nächste Balken bildet die Einkommensklasse 5000 €/Jahr ab, in der schon etwas weniger Arbeitnehmer fallen usw.

Nun gehen wir einfach mal von den 3500 € monatliches Durchschnittseinkommen aus, was somit einem jährlichen Einkommen von 42000 €/Jahr entspricht und markieren den korrespondierenden Balken mit der Einkommensklasse, in die dieser Wert fällt grün.

Na, siehst Du schon was?

Richtig! Es gibt deutlich mehr ArbeitnehmerInnen die unter diesem Durchschnittseinkommen liegen, als welche, die darüber liegen.

Die Aussage, dass das Durchschnittseinkommen bei 3500 €/Monat liegt ist rein statistisch also sicherlich nicht falsch, aber die Konklusion; den Bürgern würde es doch gut gehen unterstreicht diese Größe aus meiner Sicht ganz und gar nicht!



# Methode I – Gleitender Mittelwert und Maximum-Detektion

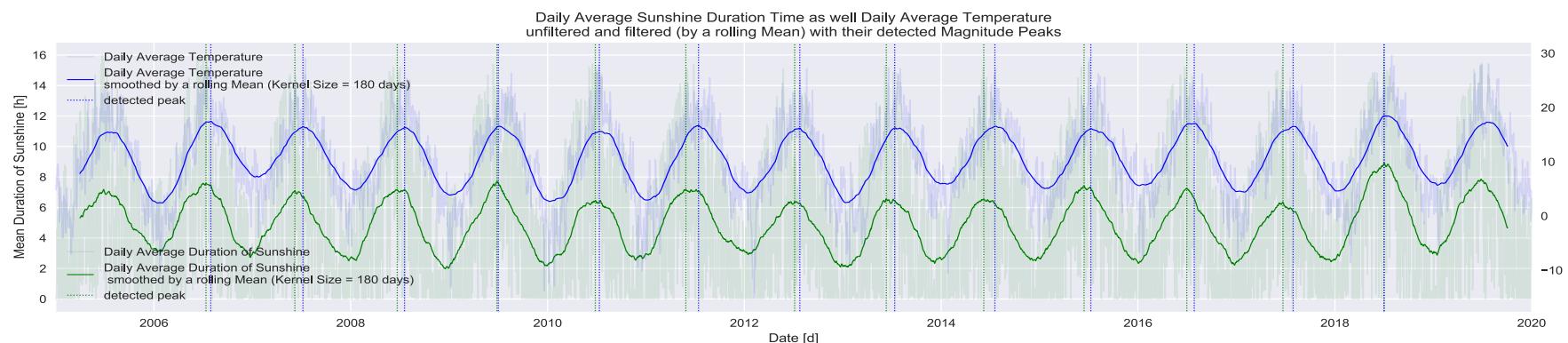
Nun geht's auch schon los mit der ersten Analysemethode.

Schon im ersten Teil haben wir den Filter „Gleitender Mittelwert“ kennen gelernt und gesehen, dass er sehr gute Dienste hinsichtlich des Glättens von streuenden Daten leisten kann. Es liegt also nahe diesen Filter als Vorprozessierung zu nutzen, um auf diese geglätteten Daten dann die Magnituden-Maxima zu finden, welche ja unsere signifikanten Tage der maximalsten Messwerte abbilden sollen. Ohne Filterung wird es nämlich sehr schwer, solche zu finden, da die ungefilterten Daten wie wild umherstreuen.

Dieser Filter arbeitet mit einem Parameter den wir von nun an  $k$  nennen wollen. Es handelt sich dabei um die bereits bekannte Fenstergröße. Man könnte stumpf behaupten, je größer  $k$  ist, umso glatter ist das Ausgabesignal. Tatsächlich ist das aber nicht so. Zumindest werden wir das bald noch sehen. Angenommen wir haben nun den optimalen Wert für  $k$ , dann können wir also die Magnituden-Maxima finden lassen. Hierzu nutzen wir einen Algorithmus zur Detektion von Signal-Extrema, wobei dieser dann nur die Maxima suchen soll. Die Minima interessieren uns nicht, da wir schließlich eh nur die Tage der Maximalwerte untersuchen wollen.

Die folgende Abbildung zeigt unser schon bekanntes Bild mit den beiden Magnituden, aber nun zusätzlich mit den detektierten Maxima visualisiert als vertikale Linien. Jede vertikale Linie verläuft also durch den detektierten maximalsten Punkt. Man kann sehen, dass der Abstand für die detektierten Maxima recht unterschiedlich ist. Es gibt also eine Streuung in den Ergebnissen. Auffällig ist aber auch in dieser Abbildung, welche nun einen noch größeren Zeitbereich repräsentiert, dass die meisten grünen Linien (bis auf die „jüngste“) alle links neben der nächsten blauen liegen.

Zudem können wir auch sehen, dass insbesondere bei der grünen Kurve das Signal (noch) nicht besonders glatt ist. Wir schieben das jetzt der Filtergüte in die Schuhe und behaupten, dass der Parameter  $k$  einfach noch nicht optimal ist. Würde nach dieser Behauptung  $k$  also ein „besserer“ Wert sein, dann würden wir doch eine geringere Streuung der Maxima um einen mittleren Wert haben, oder? Und genau diese Behauptung nutzen wir nun als ein Optimierungskriterium um  $k$  zu optimieren.



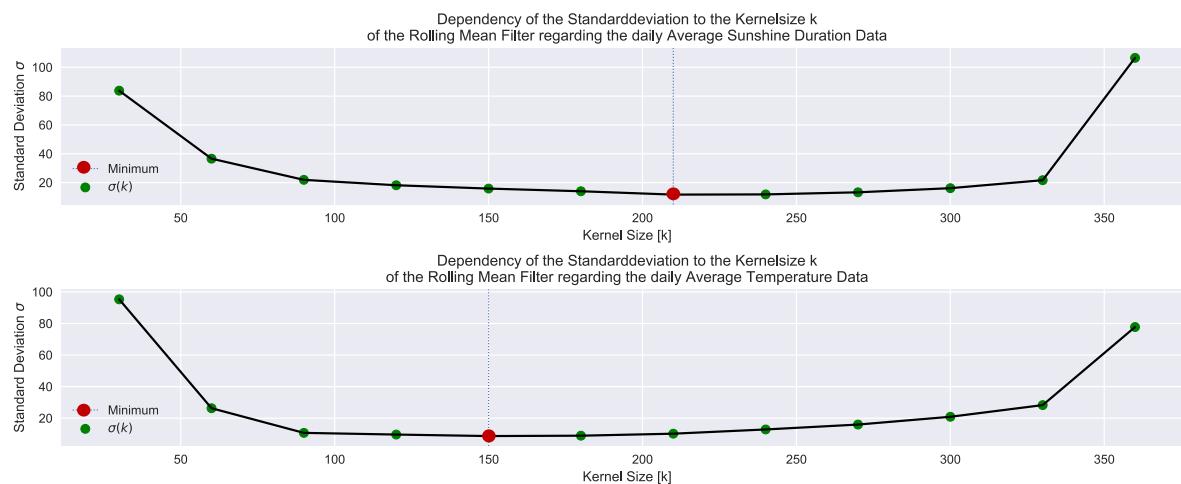
**Abbildung 5:** Überlagerungs-Plot der beiden Parameter mittlere Tagestemperatur (blaue Kurve) und tägliche Sonnenscheindauer (grüne Kurve). Die vertikalen Linien zeigen die detektierten Magnituden-Maxima. Zudem ist das unfilterte Signal ebenfalls abgebildet. Es ist zu erkennen, dass das erkennen der Maxima ohne vorhergehende Filterung sehr schwer bis unmöglich wäre.



# Methode I – Gleitender Mittelwert und Maximum-Detektion

Wir verändern den Wert  $k$  nun einfach solange, bis die Schwankung bei den detektierten Maxima kleinstmöglich ist. Dazu messen wir diese Schwankung als Standardabweichung  $\sigma$  aller detektierten Zeitpunkte des Magnitudenmaximums in Abhängigkeit von  $k$ . Das ist ein leicht zu lösendes Minimierungsproblem zur Parameteroptimierung. Wir wenden einen einfachen Gridsearch-Ansatz an. Das bedeutet, wir probieren einfach verschiedene Werte für  $k$  und schauen uns an, wie stark die Standardabweichung der ganzen detektierten Maxima-Tage für dieses  $k$  ist. Für den Wert  $k$ , bei der sie am kleinsten ist, behaupten wir einfach, dass er der „beste“ Wert ist. Diese „Brechstangen-Methode“ wird in der Informatik auch Brute-Forcing genannt und sollte aber auch nur in geringdimensionalen Parameterräumen genutzt werden!

Hier steht nämlich der Programmieraufwand in einem (noch!) guten Verhältnis zur eigentlich recht hohen Laufzeitkomplexität unseres Programms (ich nenne es dann gerne „wenig menschlicher Overhead“). Wenn wir aber mehrere Parameter zu schätzen hätten, und der Parameterraum also höherdimensional wäre, wäre das keine wirklich kluge Methode und wir würden uns eher für andere Verfahren wie der Gradientenabstiegsmethode oder ähnliche entscheiden.



**Abbildung 5:** Ergebnisse der Gridsearch basierten Parameteroptimierung für die Fensterlänge  $k$  zum Glätten des Signals anhand eines rollenden Mittelwerts. Die vertikalen Linien und roten Punkte zeigen das lokale Minimum der Funktion im Untersuchungsbereich an.

In der nebenstehenden Abbildung sieht man das Ergebnis der Parameteroptimierung anhand der Minimierung der Standardabweichung der detektierten Magnitudenmaxima-Tage, für beide Wetterparameter.

Interessant ist, dass für beide Wetterparameter unterschiedliche Optimalwerte für  $k$  existieren.

Zudem nimmt die Streuung ab dem Optimalwert wieder zu. Dieses Verhalten einer zu minimierenden Zielfunktion ist oft der Fall und wichtig für den Optimierungsvorgang. Denn ihre Minimumsstelle bildet den optimalen Parameterwert ab. Wenn wir sie nicht finden, weil wir an der falschen Stelle suchen oder sie gar nicht existiert, haben wir ein Problem...

Zur Detektion des Tages der maximalen Sonnenscheindauer ist der optimale Wert für  $k$  210, und für die Detektion des Tages der maximalen Temperatur 150.

Falls Du Dich nun noch fragst, warum der Wert der Standardabweichung also die Streuung mit steigenden  $k$  nicht immer besser sondern irgendwann wieder schlechter wird; Der Filter fängt irgendwann an das Signal zu übersteuern. Das hat eventuell mit dem Sinusartigen Verhalten des Signals zu tun, da es bei bestimmten Fensterlängen als vielfache der Periodenlänge vermutlich zu Interferenzen kommt.



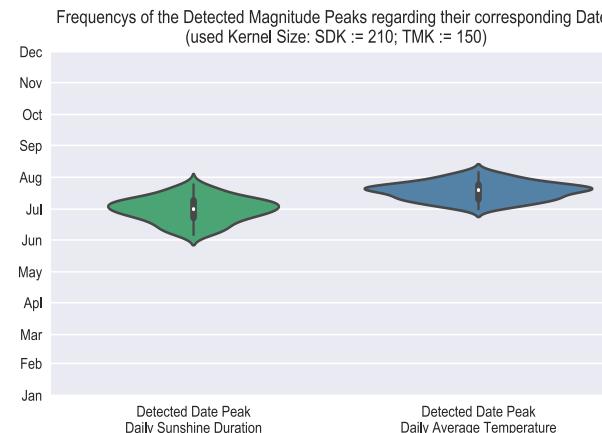
# Methode I – Gleitender Mittelwert und Maximum-Detektion

Nachdem der Parameter  $k$  für beide Signale optimiert wurde, werden nun die Statistiken ermittelt, um unser Model zu prüfen. Dabei prüfen wir jeweils für jedes Signal den mittleren Zeitwert zum Maximum und die Standardabweichung (...die wir ja eigentlich durch die  $k$ -Optimierung schon längst kennen) der ermittelten Werte.

Als Ergebnis für die maximalst tägliche Sonnenscheindauer wird der Tag auf den 29. Juni geschätzt, wobei eine Standardabweichung von 11.7 Tagen als Streuung beachtet werden muss. Bezüglich des Tages mit der maximalen mittleren Tagestemperatur schätzt das Model den 17 Juli mit einer zugehörigen Standardabweichung von 8.6 Tagen. Bedenkt man, dass der 21. Juni bekannter Maßen der längste Tag ist, ist der 29. Juni als Schätzung schonmal gar nicht sooo schlecht. Aber die Standardabweichung von 11 Tagen ist eher ziemlich hoch. Auch bei der mittleren Tagestemperatur liegen wir in einem Bereich, der vielleicht ganz gut geschätzt wurde, aber immer noch mit einer hohen Unsicherheit der Schätzung verbunden ist.

$k$	$\sigma_{SDK}$	$\sigma_{TMK}$	$\bar{x}_{SDK}$	$\bar{x}_{TMK}$
30	83.78090208	95.33686425	182.4711538	172.5641026
60	36.61501317	26.30891883	181.3544304	200.3188406
90	21.93802999	10.68576976	174.7230769	200.7536232
120	18.18875392	9.62373791	176	198.7538462
150	15.88121588	8.665538225	179.3076923	199.453125
180	14.02452337	8.900038953	178.515625	200.234375
210	11.69454783	10.17272702	181.4571429	199.875
240	11.83812629	12.8549512	176.9104478	199.4848485
270	13.29225488	15.95759026	176.3088235	200.1641791
300	16.15557811	20.89024062	170.7906977	195.4313725
330	21.69131107	28.28743341	171.6764706	195.4333333
360	106.530327	77.77083295	216.2592593	207.0434783

**Tabelle 1:** Ergebnisse der Gridsearch basierten Parameteroptimierung zur Fensterlänge  $k$  in Form des: Mittelwerts  $\bar{x}_{SDK}$  des vom Model erkannten jährlichen Maximalwerts bzgl. der täglichen Sonnenscheindauer, des Mittelwerts  $\bar{x}_{TMK}$  vom erkannten jährlichen Maximalwert bzgl. der mittleren Tagestemperatur, sowie die zugehörigen Standardabweichungen  $\sigma_{SDK}$  und  $\sigma_{TMK}$ . Die Tage sind als Tages-Index angegeben. D.h. 1 entspricht dem 1. Jan und 365 dem 31. Dez.



**Abbildung 6:** Violinen Plot zur Visualisierung der vom Model detektierten Maximum-Tage seitens Model-behafteter Abweichungen.

Die Tabelle zeigt alle Werte der Optimierung nochmal. In rot sind die jeweils optimalen Werte hervorgehoben.

Zudem habe ich die Ergebnisse noch als Violinen-Plots visualisiert.

Dadurch bekommt man einen besseren Eindruck, wie stark die Modellergebnisse jedes detektierten Maximums um ihren Median (weißer Punkt) streuen.

Es zeigen sich für beide Wetterparameter annähernd gleichmäßige Abweichungen um den Median. Das sieht man an dieser „Rochen-artigen“ Form beider Figuren.

Sie ist vor allem bei der grünen Figur Gauß-Kurven-artig. Gauß-Kurven-artige Verteilungen sind ein gutes visuelles Indiz für eine mögliche Normalverteilung, welche wiederum - subjektiv - für sinnvolle Ergebnisse sprechen könnte. Wobei das Kontext-Abhängig ist, da es auch andere „natürliche“ Verteilungsformen gibt.

Die blaue Figur ist aber mit strengem Auge betrachtet etwas weniger symmetrisch. Sie scheint ganz leicht schief zu sein.

Das ist aber auch nicht schlimm, da schiefe Verteilungen ebenso natürlich sein können.



# Methode I – Gleitender Mittelwert und Maximum-Detektion

Die wohl am leichtesten zu interpretierenden statistischen Diagramme sind Histogramme. Man kennt sie schon aus der Schule und den vielen interessanten und manchmal teils fragwürdigen Statistiken aus den Medien.

Sie bilden Häufigkeiten von Beobachtungswerten auf bestimmte Klassen ab. In Zusatzabbildung i hatten wir ja bereits ein sehr interessantes Histogramm bzgl. der Einkommensklassen und ihrer Häufigkeiten hinsichtlich sozialversicherter Arbeitnehmer gesehen.

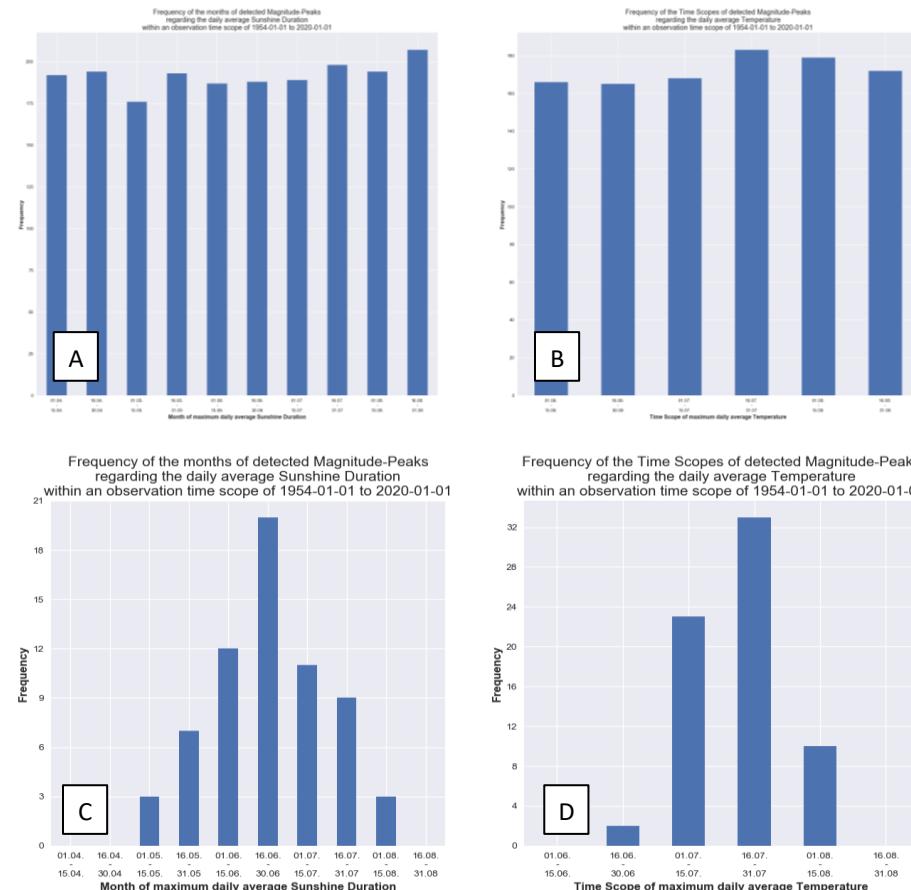
Deshalb habe ich die Ergebnisse unseres Models nochmal als Histogrammplots dargestellt (siehe Abbildung rechts).

Jeder Balken steht für einen Zeitbereich (in diesem Fall konkret ein halber Monat). Die Länge der Balken gibt die Häufigkeit an, wie oft ein Maximum vom Maximum-Detektor für diesen Zeitbereich erkannt wurde. Bevor gefiltert wurde, werden für jeden halben Monat annähernd gleich viele Maxima detektiert (siehe A und B). Wir gehen hierbei von einer Gleichverteilung ihrer Häufigkeiten aus.

Nachdem die Daten geglättet wurden, detektiert der Algorithmus weit weniger Maxima und deren Verteilung scheint jetzt nicht mehr gleichverteilt zu sein siehe (C und D). Man kann klar signifikante Häufigkeiten der Maxima in den relevanten Monatsbereichen sehen, in denen wir sie auch erwarten würden.

Die Histogramme unterstreichen also nochmal unsere Modellgüte. Und wenn wir gar nicht so sehr interessiert wären, den Tag der maximalen Magnitude zu schätzen, sondern uns mit einem größeren Zeitbereich begnügen, dann wären diese Analyseergebnisse bereits schon ausreichend genug.

In D kannst Du übrigens auch die nicht ganz symmetrische Verteilung sehen, wie schon in Abbildung 6. dafür angenommen. Der Balken direkt links neben dem Längsten ist weitaus länger als der rechts neben dem längsten Balken. Für C sieht das Ganze verglichen eher symmetrischer aus.



**Abbildung 7:**  
Histogramme der Maximalwerte bzgl. ungefilterter Daten (A, B) und gefilterter Daten (C, D). Es zeigt sich vor dem Behandeln eine annähernde Gleichverteilung der Maximalwerte über alle Zeitbereiche. Nach dem Behandeln sind deutliche Anhäufungen in Form langerer Balken zu sehen. Als Zeitbereich wurde nicht das ganze Jahr sondern der vom 01.04 – 31.08 definiert.

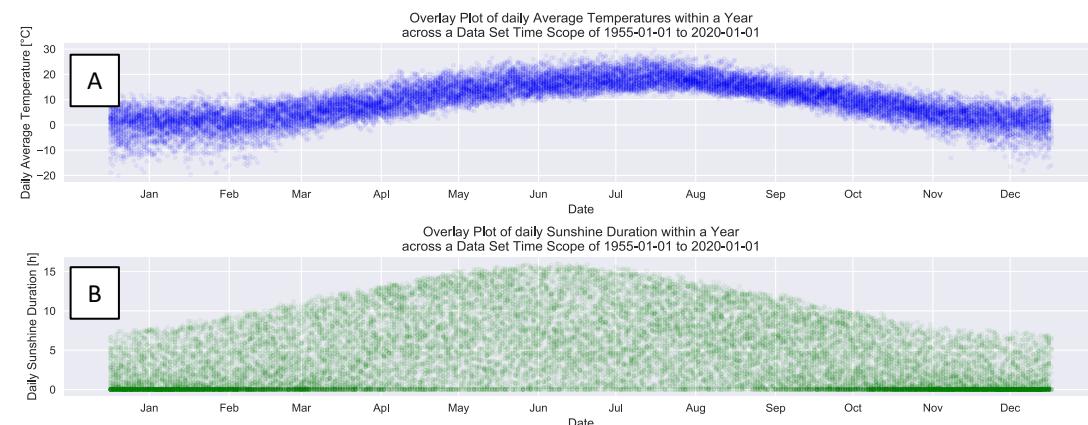


# Methode II – Anpassung einer Sinusfunktion

Die erste Methode hat Ergebnisse geliefert, welche zumindest schonmal recht plausibel erscheinen. Dennoch ist die Standardabweichung nicht gerade gering. Ich möchte das Modell nicht schlecht reden... aber so richtig schön ist es auch nicht. Die Datenglättung mittels gleitenden Mittelwert plus anschließender Maximum-Detektion ist wohl noch nicht das Ultimo, selbst mit einer Parameteroptimierung nicht.

Mit Methode II wollen wir es nun noch besser machen. Der Ansatz ist dabei ein vollkommen anderer. Als Basis dient uns die Annahme, dass das Signal ein Sinussignal ist. Das zeitliche Verhalten unserer beiden Wetterparameter soll sich exakt wie eine Sinusfunktion verhalten. Ein Sinus-artiges Signal zeichnet sich durch eine zeitlich konstante Periodenlänge ab. In unserem Kontext ist das genau ein Jahr bzw. 365 Tage. Im Prinzip spielt sich also alles innerhalb dieser Periode ab. Und deshalb betrachten wir unsere Daten jetzt gar nicht mehr als ganze Zeitreihe sondern nur noch innerhalb dieser Periode.

Um das Ganze zu verdeutlichen schauen wir uns dazu einfach die folgende Abbildung an. Diese betrachtet jetzt nur noch die Zeitachse auf einer Länge von 365 Tage und die Magnitude weiterhin auf der Y-Achse. Die Magnitudenwerte des gesamten Zeitraums wurden jetzt also auf den Periodenzeitraum abgebildet. Das bedeutet, dass die Datenpunkte nun aber auch bzgl. der Zeitachse übereinander liegen. Jeder Datenpunkt vom 1. Januar egal zu welchem Jahr liegt nun also auf dem ersten Periodenzeitzpunkt, jeder des 10. Januars auf dem 10. Periodenzeitzpunkt und jeder des 31. Dezembers auf dem 365. usw.



**Abbildung 8:** Visualisierung aller Datenpunkte innerhalb des Periodenzeitraumes  $T = 365$ . Das sinusartige Verhalten zeigt sich insbesondere bei der mittleren täglichen Temperatur (A) weiterhin sehr gut. Bei der täglichen Sonnenscheindauer (B) ergibt sich aber ein eigenartiges Bild. Während die Temperatur um einen häufigen Wert streut, scheint die Maximaltemperatur eine ganz anderes Verteilungsmuster aufzuweisen. Auf dem ersten Blick ist es sehr homogen. Im unteren Magnitudenbereich zeigen sich aber Anhäufungen in den kälteren Monaten.

Bei der Temperatur kann man nun weiterhin das Verhalten als Sinus sehen. Die Punkte scheinen um einen relativ häufigen Wert geringfügig zu streuen. Es soll für uns keine Rolle spielen, aber ich finde interessant, dass die Streuung auch Zeitabhängig ist. Sie ist im September rum am geringsten und in den Wintermonaten am größten. Aber schau selber nach. Erkennst Du es?

Bei der effektiven Sonnenscheindauer sieht das alles allerdings etwas anders aus. Im niederen Magnitudenbereich gibt es im Bereich der kälteren Monate Anhäufungen → Klar in den kälteren Monaten ist der Himmel häufiger bedeckt ergo ist die gemessene effektive Sonnenscheindauer kürzer und in den Daten sogar häufig mit 0 h beziffert. Ansonsten scheint es eine relativ homogene Verteilung der Punkte zu geben.

Ein weiteres sehr interessantes Muster ist die sehr harte Grenze im Maxima-Wertebereich über den gesamten Zeitbereich.

Aber auch das kann man logisch erklären. Denn auch wenn die effektive Sonnenscheindauer vom Wolkenbedeckungsgrad gemindert werden kann. Es gibt eben auch viele unbewölkte Tage an denen die Sonne ihr volles Zeitpotenzial ausschöpfen kann... nur ist das aber eben hart begrenzt. Das Ganze hat also astronomische Gründe. Die Erde-Sonne-Kinematik bestimmt hart die Tageslängen. Sie können effektiv nur durch Bewölkung verkürzt, aber eben nicht verlängert werden.



# Methode II – Anpassung einer Sinusfunktion

Es geht nun darum, eine Sinusfunktion passend zu jedem Wetterparameter anhand der jeweiligen Daten der jeweiligen Wetterparameter als Model zu finden. Anhand dieser Sinusfunktion können wir dann hoffentlich noch exakter den Maximalwert bzw. vielmehr den Zeitpunkt des Maximalwertes schätzen. Wir nutzen dazu das Prinzip der [Regressionsanalyse](#)<sup>1</sup> und passen drei wichtige Parameter dieser Sinusfunktion an. Wir nutzen jetzt echtes maschinelles Lernen und trainieren ein Model an unsere Messdaten.

Die folgende Funktion, beschrieben durch Formel 1, sei unsere Sinusfunktion, welche unser Model repräsentiert. ...nun wird's also doch etwas mathematischer...

$Y(t)$  ist die Magnitude zum Zeitpunkt  $t$ , also der vom Model geschätzte Messwert korrespondieren zum jeweiligen Wetterparameter.  $Y_{max}$  ist die Amplitude, also der Maximalwert der Magnitude, den das Model schätzen soll.  $Y_{shift}$  ist ein Magnitudenversatz bzgl. der Y-Achse. Ein „echter“ Sinus schwingt nämlich gleichmäßig um den Nullwert bzw. der Y-Achse → also bei  $y = 0$ , aber bei unseren Daten ist das eben nicht der Fall, da der durchschnittliche minimale Magnitudenwert etwa im Fall der Temperaturzeitreihe bei so ca. 0°C liegt, dafür aber der durchschnittliche maximalste bei etwa 19°C. Um dies zu korrigieren, verschieben wir die Werte durch diesen Parameter so, dass alles wieder „virtuell“ gleichmäßig um den Null-Punkt schwingt.  $T$  ist die Periodendauer ( $T = 365$  Tage := ein ganzes Jahr).  $t$  ist – wie schon erwähnt – der Periodenzeitzpunkt, korrespondiert also zum Datum, zu dem das Model den Wetterparameterwert schätzen soll.  $\varphi$  ist der sogenannte Phasenwinkel. Das ist ein wichtiger Parameter, der die Magnitude entlang der X-Achse verschiebt.  $\pi$  sollte sicherlich jedem bekannt sein. Das ist natürlich die äußerst faszinierende Kreiszahl, die irgendwie fast immer und überall in der Mathematik aufzutauchen scheint...selbst da wo ich sie irgendwie gar nicht vermute. Hier beim Sinus würde ich sie natürlich definitiv vermuten!

Die 3 einzigen Parameter die es zu schätzen gilt, sind hierbei nun  $Y_{max}$ ,  $Y_{shift}$  und  $\varphi$ . Diese werden bei unseren beiden zu analysierenden Wetterparametern jeweils unterschiedlich sein. Sie charakterisieren quasi diese beiden Wetterparameter, wenn man so will.

$$Y(t) := Y_{max} \cdot \sin\left(\frac{2 \cdot \pi}{T} \cdot t + \varphi\right) + Y_{shift} \quad (1)$$

Doch bevor wir uns jetzt ran machen und die Parameter der beiden Sinusfunktionen zu den Wetterparametern optimieren (lassen), müssen wir noch die Daten bzgl. der täglichen Sonnenscheindauer etwas filtern. Denn an die können wir mit Sicherheit noch keine Sinusfunktion sinnvoll anpassen.

Dazu werden wir bestimmte Datenpunkte aussieben. Wir verwenden zwei Filter. Zum einen wollen wir nur die Maximalwerte einer Wochenzusammenfassung haben. Damit werden schonmal eine ganze Menge Punkte mit sehr geringer Magnitude wegfallen. Vor allem die Anhäufungen im unteren Bereich. Denn sonst würde unser darauf aufsitzende zweiter Filter nämlich nicht richtig funktionieren. Wir werden nämlich einen DBSCAN-Algorithmus nutzen, um Anhäufungen zu detektieren und valide Werte „durch zu lassen“. Noch mehr maschinelles Lernen also. Denn der DBSCAN-Algorithmus ist einer der bekanntesten Algorithmen im Bereich des Maschinellen Lernens, mit dem man tausende andere Anwendungen machen kann. Unter anderem wird er zur Detektion von Kreditkarten-Betrug ([guck mal hier](#)<sup>2</sup> dazu das Paper von BHARATI ET AL. an) oder der Analyse des Kaufverhaltens von Kunden anhand ihres Surfverhaltens angewendet. Du fragst Dich manchmal, warum du relativ passende Werbung beim Surfen im WWW bekommst? Er könnte der Grund dafür sein...

Die nächste Abbildung wird die Ergebnisse nach dem Filter zeigen und alles wesentlich verständlicher machen.

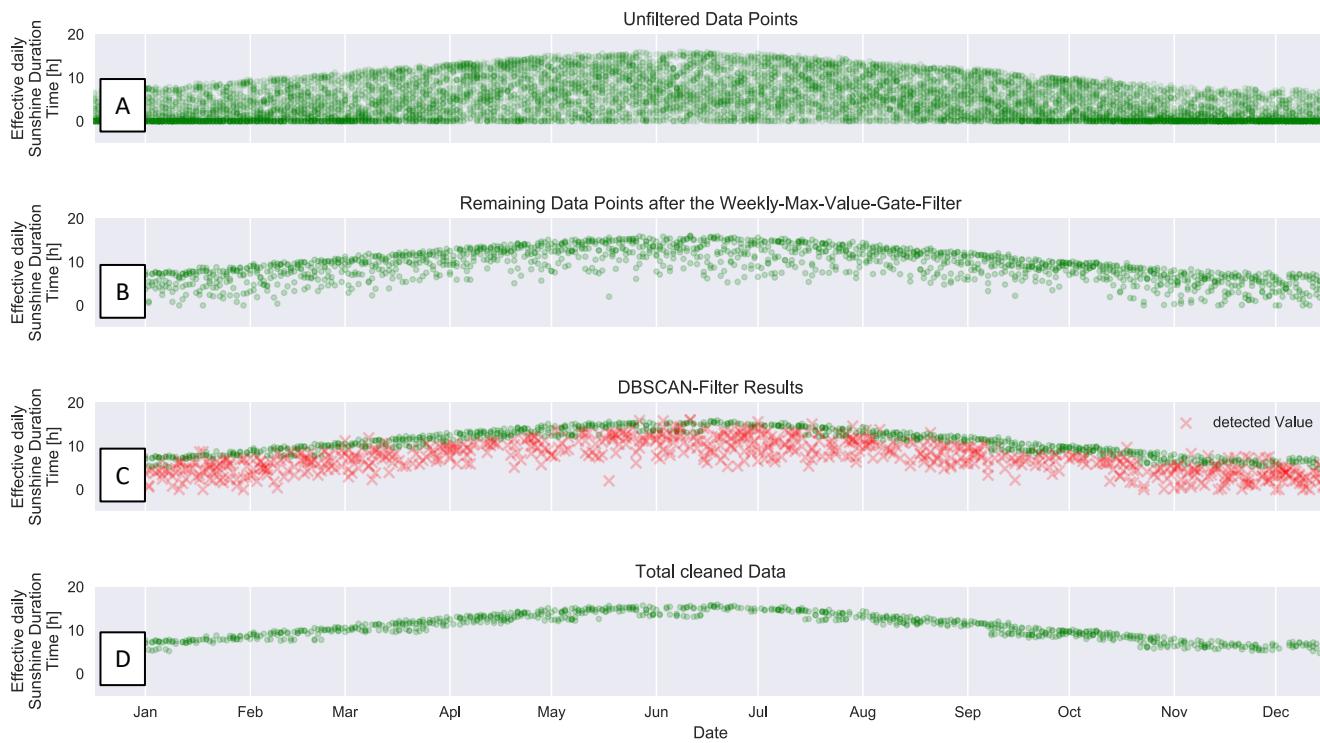
1: <https://de.wikipedia.org/wiki/Regressionsanalyse>

2: <https://www.ijsr.net/archive/v7i10/ART20192099.pdf>



# Methode II – Anpassung einer Sinusfunktion

In Abbildung 8 sehen wir den Ergebnisverlauf der Datenfilterung. A Zeigt alle Datenpunkte ohne Filterung noch einmal. In B ist das Ergebnis nach dem Filtern der wöchentlichen Maximalmagnituden zu sehen. Man kann sehen, dass nun aus dem unteren Bereich schon sehr viele Punkte weg gefallen sind.



**Abbildung 9:** Ergebnisse der Datenfilterung.

Es werden zwei Filterstufen plus der Vorher- und Nachherzustand gezeigt. Wobei A den Vorherzustand also die ungefilterten Datenpunkte zeigt. Anschließend werden die wöchentlichen Maximalwerte gefiltert (B). Seriell im Anschluss an diesem Filter arbeitet ein DBSCAN-basierter Filter und filtert alle noch vorhandenen Datenpunkte innerhalb einer bestimmten Punktwolkendichte (C). Nach dem gesamten Filterprozess sind fast nur noch die maximalsten Datenpunkte des gesamten Datensatzes übrig (D). Ein mehr täglicher Maximalwertfilter hätte nur sehr wenige Datenpunkte gefiltert. Insbesondere auch im Niedertemperaturbereich wären noch viele, somit ungünstige, Daten übrig. Durch den wöchentlichen Filter haben wir noch viele Datenpunkte zur Verfügung. Bei einem monatlichen Filter wären sehr viele Datenpunkte verloren gegangen. Durch diese Kombination an Filtern, wurden also immer noch genug Datenpunkte beibehalten, um eine vertrauensvolle Datenschätzung anhand der nun anschließenden Sinusanpassung vornehmen zu können.

Interessant ist, dass auch diesmal wieder aus einer anfänglichen Gleichverteilung (homogene Punktwolkendichte in den oberen Magnitudenbereichen) plötzlich ein anderes Verteilungsmuster mit einem klaren Gradient bzgl. der Punktwolkendichte zu sehen ist.

Um nun aber eine wirklich gute Grundlage für die Parameterschätzung zur Anpassung der Sinusfunktion zu erhalten, filtern wir noch mit dem DBSCAN-Algorithmus. Dazu nutzen wir die nun inhomogene Punktwolkendichte aus, und lassen den DBSCAN-Filter alles aussieben, dass unterhalb einer bestimmten Punktwolkendichte liegt.

Diesen Dichteparameter müssen wir in dem DBSCAN-Algorithmus noch zusammen mit einem weiteren Parameter einstellen. Wir müssen also wieder Parameter optimieren. Aber dass können wir einfach optisch durch die Visualisierung „per Hand“ machen. Wir probieren solange rum, bis es für optisch gute Ergebnisse sorgt.

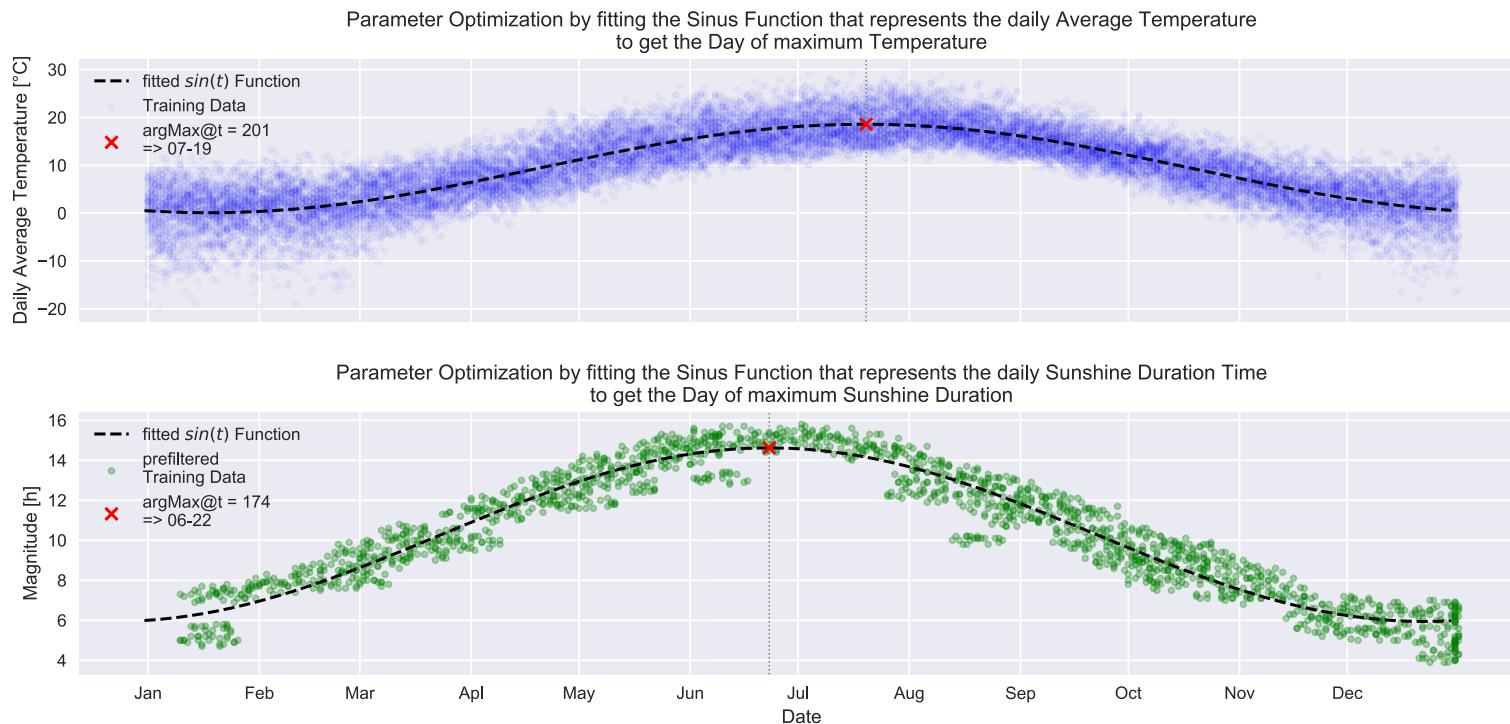
Und man kann schließlich anhand der rot markierten Datenpunkte sehen, dass der Filter gut arbeitet und fast nur noch die maximalsten Datenpunkte durch lässt (siehe Teil D).

Das Ergebnis ist sehr zufrieden stellend und auf die so gefilterten Datenpunkte sollte sich nun ohne Probleme eine passende Sinusfunktion bzw. deren Parameter finden lassen.



# Methode II – Anpassung einer Sinusfunktion

Auf Basis der vorgefilterten Daten werden wir nun die Parameterschätzung vornehmen und für jeden Wetterparameter seine eigene Sinusfunktion anpassen. Mit dieser können wir dann die Maximalwerte bzw. deren Zeitpunkte hoffentlich sehr gut abschätzen. Die folgende Abbildung zeigt die Ergebnisse der Anpassung und die Maximalwerte.



**Abbildung 10:** Ergebnisse der Parameterschätzung zur Anpassung der Sinusfunktion sowie der dadurch ermittelten Maximalstellen. Die schwarzen gestrichelten Kurven zeigen jeweils die angepasste Sinusfunktion. Der Rote Punkt markiert in beiden Fällen die Maximalstelle der Sinusfunktion. Der vertikale Strich lässt eine bessere Beurteilung des korrespondierenden Zeitpunktes zu. Die Anhäufungen zum Jahresende sind Filterartefakte die beim Filtern der wöchentlichen Maximaltemperaturen auftraten. Man könnte sie beseitigen, aber sie spielen bei dieser Modellierung keine Rolle da sie bzgl. der Magnitude gleichverteilt auftreten.

Die Ergebnisse nach der Parameterschätzung können sich wirklich (schonmal) sehr gut sehen lassen. Die schwarzen gestrichelten Kurven repräsentieren beide die jeweils an die Wetterdaten angepassten Sinusfunktionen.

Anhand dieser können dann die Maximalwerte (rote Kreuze) ermittelt werden.

Selbst wenn im nahen Zeitbereich um die Maximalzeitpunkt herum nie Datenpunkte existiert hätten, könnten wir die Maximalwerte dennoch sehr gut durch diese Methode schätzen.

Eine Glättung könnte das in so einem Fall übrigens nicht, da sie von den näheren Nachbarpunkten zum Erwartungspunkt abhängt.

Der vom Modell geschätzte Tag für die maximale Sonnenscheindauer ist der 22. Jun.

Und der für den Tag der maximalsten Temperatur der 19. Juli.



# Methode II – Anpassung einer Sinusfunktion

Vorab sind diese Ergebnisse doch schonmal richtig klasse, bedenkt man, dass ja der tatsächliche Tag mit der längsten Sonnenscheindauer der 21. Juni ist und das Vorab-Modell den 22. Juni anhand der Daten mit demnach gerade mal einen Tag Unterschied schätzt.

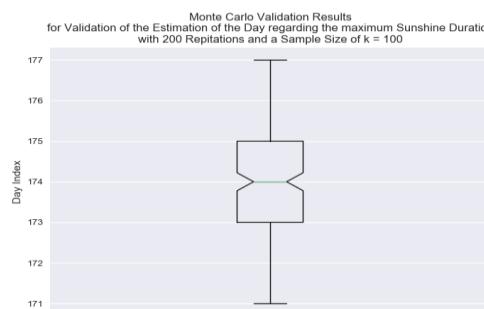
Aaaaaaber dieses Modell ist noch kein gutes, solange wir nicht auch ein wenig kritisch bzw. statistisch drauf geschaut haben.

Denn bisher haben wir einfach alle Datenpunkte genommen und die Sinusfunktion daran angepasst. Das ist kein ganz so gutes Vorgehen, wenn man es nur dabei belässt. Wir haben zwar einen gewissen positiven Trend durch das Modell erkennen können, dass es gut funktionieren könnte, aber kennen überhaupt nicht dessen Unsicherheit. Klar, mit dem Erwartungswert zumindest bei der Sonnenscheindauer haut es ja eigentlich super hin. Aber der Erwartungswert ist als Vergleich nicht immer bekannt.

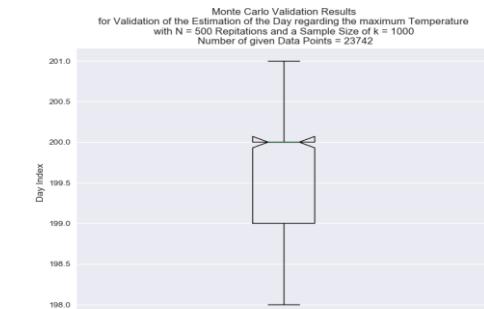
Deshalb muss ein Model immer validiert werden!

Validieren wir also unser Modell noch. Denn da wir sehr viele Datenpunkte haben, können wir mit einem kleinen Trick die Modelgüte sehr gut bestimmen. Wir nutzen einen sogenannten Monte-Carlo-Kreuz-Validierungs-Ansatz und passen dazu unsere Sinusfunktionen an verschiedenen Stichproben unserer Daten an. D.h. wir nehmen nun also gar nicht alle Datenpunkte, sondern einfach nur einen bestimmten Teilmengen davon. Dazu ziehen wir zufällig Datenpunkte (ohne Zurücklegen) aus dem Datensatz und passen daran die Sinusfunktion an, um die Zeitpunkte der Tage des Maximalwertes zu schätzen. Das machen wir einige Male mit also jeweils anderen Stichproben aus zuvor zufällig gezogenen Datenpunkten. Der Clou des Ganzen ist nun, dass wir einen Mittelwert und die Varianz des Mittelwerts anhand vieler Proben des Modells schätzen können. Anhand dieser Varianz des Mittelwerts können wir auf die Unsicherheit des Modells schließen. Je geringer sie ist, desto sicherer können wir uns unserem Model sein. Das wird dann noch wichtig, wenn wir unsere Hypothese überprüfen.

Uns interessiert vor allem das Modell, mit dem wir die mittlere Tagestemperatur im Jahres-Maximum schätzen. Es ergibt sich ein Modellmittelwert von  $\bar{x}_{sin} = 199.596$  und der Median liegt bei 200. Die Standardabweichung des Mittelwerts beträgt 1.018



**Abbildung 11:** Ergebnisse der Monte-Carlo-Analyse bzgl. der mittleren Sonnenscheindauer und deren Zeitpunktes des maximalen Jahreswertes. Der Median des Modellmittelwerts liegt bei 174 was dem 22. Juni entspricht. Die Standardabweichung beläuft sich auf ca. einen Tag.



**Abbildung 12:** Ergebnisse der Monte-Carlo-Analyse bzgl. der mittleren Tagestemperatur und deren Zeitpunktes des maximalen Jahreswertes. Der Median des Modellmittelwerts liegt bei 200. was dem 18. Juli entspricht. Auch hier beträgt die Standardabweichung einen Tag wobei sie nur linksseitig gilt.



# Methode II – Anpassung einer Sinusfunktion

Nun können wir nach unserer Validierung wesentlich objektiver auf die Modellgüte unseres Ansatzes schauen.

Die Monta-Carlo-Kreuz-Validierung sagt uns nun weiterhin den 22. Juni als wahrscheinlichsten Tag für die maximalst Sonnenscheindauer zu und verglichen mit dem 21. Juni als Tag der „astronomisch“ längsten Sonnenscheindauer ist das ein sehr plausibles Ergebnis. Gemessen an der Standardabweichung von ca. einem Tag ist die Unsicherheit des Models bzgl. der Schätzung des Tages mit der längsten effektiven Sonnenscheindauer zudem wirklich super klein. D.h. auch die Modellunsicherheit, gemessen an dieser Standardabweichung, ist wirklich sehr gut – da sehr gering!

Beim Model zur Schätzung des Tages der jährlichen Maximaltemperatur schätzt das Modell den 18. Juli und hat eine Unsicherheit von ca. einem Tag gemessen an der Standardabweichung, wobei diese nicht beidseitig gilt, sondern nur rechtsseitig, also positiv zum Median ist.



# Methode III – LISSAJOUS-Ansatz

Jetzt kommen wir zum letzten Modellierungsansatz. Meinem persönlichen Favoriten!

Es ist ein sehr „exotischer“ und wirklich sehr eigenartiger Ansatz, mit dem Gedanken einer möglichen Problemvereinfachung durch Transformation. Viele mathematische Probleme lassen sich durch Transformationen von mathematischen „Objekten“ aus ihren Basis-Räumen in andere Räume vereinfachen. Deshalb gibt es etwa Konzepte wie: FOURIER-, LAPLACE-, LORENZ- oder HILBERT-Transformation - um nur einige Wenige, aber wohl sehr Bekannte zu nennen. Wenn Du nicht weißt, was eine solche Transformation ist, lass es einfach so stehen und kümmere Dich nicht weiter darum. Andernfalls kannst Du einfach mal bei WIKIPEDIA [hier](#)<sup>1</sup> vorbei schauen...aber ich warne Dich vorher schonmal vorab ;).

Aber worum geht es denn nun eigentlich?

Dass unsere Wetterdaten (in guter Näherung) als Sinusfunktionen repräsentiert werden können, haben wir ja bereits klar sehen und zeigen können. Nun machen wir uns dies nochmal zu Nutze und berechnen diesmal den Tag des jährlichen Maximalwertes seitens der mittleren Tagstemperatur in relativer Weise. Dazu betrachten wir gar nicht mehr weiter die Daten der täglichen Sonnenscheindauer, sondern nehmen den Tag des 21. Juni als eben den mit der längsten Sonnenscheindauer an. Diesem können wir jetzt eine Sinusfunktion zuordnen, welche also ihr Maximum bei  $t = 173$  hat (der 21. Juni ist ja der 173. Tag in unserer Periode  $T = 1$  Jahr). Im weiteren Verlauf sei dies nun unsere sogenannte Referenzfunktion und nennen sie auch so.

Ok, und für was? Für die Sinusfunktion der mittleren Tagstemperatur. Die kennen wir wiederum nur durch ihre Werte, also den bekannten Datenpunkten. Aber so richtig interessiert uns auch gar nicht die Sinusfunktion an sich. Also wir wollen diesmal gar nicht die ganzen Parameter schätzen. Denn was uns nun eigentlich interessiert ist nur die Phasenverschiebung (zeitlicher Versatz) dieser Sinusfunktion relativ zur Referenzfunktion. Denn alleine die Phasenverschiebung beider Sinus-Kurven plus dem 21. Juni ergibt nach diesem Modellansatz den Tag der maximalen mittleren Tagstemperatur im Jahr.

Aber wie kommt man auf die Phasenverschiebung, ohne nun direkt diese Sinusfunktion zu kennen? Wir nutzen das Prinzip von [LISSAJOUS-Figuren](#)<sup>2</sup> bzw. einer ganz bestimmten, welche sich nämlich auf zwei Sinus-Signale mit gleicher Wellenlänge bezieht. LISSAJOUS-Figuren gehen übrigens auf die Arbeiten des französischen Physikers [JULES ANTOINE LISSAJOUS](#)<sup>3</sup> zurück. Diese Figur besteht aus den Magnitudenwerten der jeweiligen beiden Sinus-Funktionen bezogen auf ihren gleichen Periodenzeipunkt abgebildet auf die beiden Achsen eines Diagramms. Zu jedem Zeitpunkt müssen wir also sowohl die Magnitude des einen Sinus als auch die des andern Sinus kennen, Und bilden nun die Magnitude des einen Sinussignals auf die X-Achse und die des anderes auf die Y-Achse ab. Der Zeitpunkt spielt nun nur noch eine implizite Rolle. Dabei entsteht je nach Phasenverschiebung eine Ellipse oder Kreis oder Strich.

Wir werden unsere Datenpunkte einfach in ein solches Diagramm Plotten und die so entstandene Ellipse bzgl. ihrer LISSAJOUS-Ellipsen-Geometrieparamter anpassen. Dazu verwenden wir einen Programm-Code, der freundlicherweise von BEN HAMMEL, & NICK SULLIVAN-MOLINA auf deren [GitHub-Repository](#)<sup>4</sup> zur freien Verwendung gestellt wurde. Wir verwenden die Referenzfunktion um zu jeden gegebenen Zeitpunkt der gegebene Magnitudenwerte eine korrespondierende Referenzmagnitude zu erzeugen.

1: [https://de.wikipedia.org/wiki/Basiswechsel\\_\(Vektorraum\)](https://de.wikipedia.org/wiki/Basiswechsel_(Vektorraum))

2: <https://de.wikipedia.org/wiki/Lissajous-Figur>

3: [https://de.wikipedia.org/wiki/Jules\\_Antoine\\_Lissajous](https://de.wikipedia.org/wiki/Jules_Antoine_Lissajous)

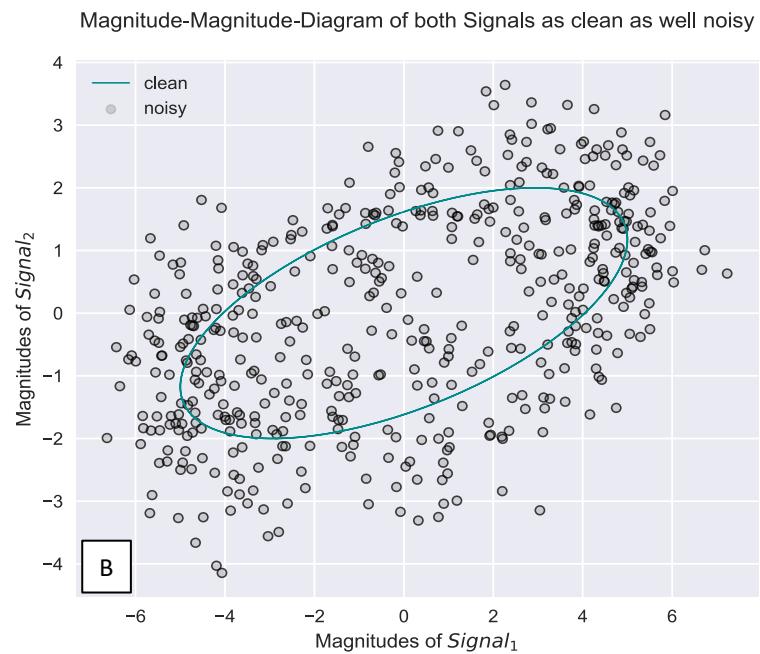
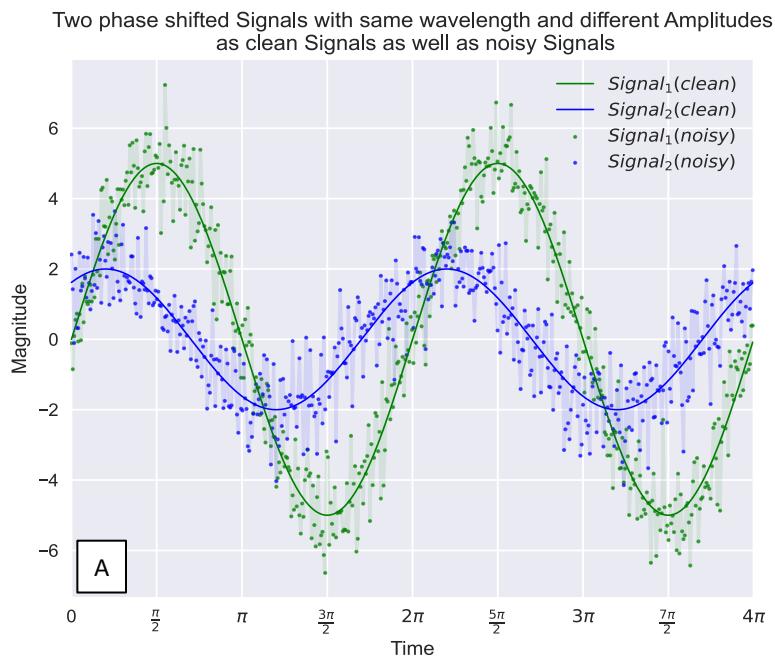
4: <https://github.com/bdhammel/least-squares-ellipse-fitting>



# Methode III – LISSAJOUS-Ansatz

Im linken Teil (A) der Abbildung 13 sehen wir zwei Sinus-artige Signale - sowohl verrauscht als auch sauber - welche zudem gegeneinander phasenverschoben sind. Wir können das wieder sehr gut anhand ihrer Extrema sehen. Etwa der gegeneinander zeitlich verschobenen Maxima. Im rechten Abbildungsteil sehen wir die dazu korrespondierende LISSAJOUS-Figur in Form einer gestreuten Punktwolke bzgl. der beiden verrauschten Signale, sowie einer perfekten Ellipse korrespondierend zu den beiden sauberen Signalen. Tatsächlich ist diese scheinbar „perfekt“ und stetige Ellipse aus genauso vielen Punkten „zusammengesetzt“ wie auch die dazugehörigen „perfekten“ Sinussignale aus einzelnen Punkten besteht. Aber da sie nicht verrauscht sind, und wir zudem sehr viele Punkte haben, scheint es, als wäre es eine stetige „perfekte“ Linie.

Die Idee ist nun, anhand einer gegebenen Punktwolke, nämlich der unserer unseren Datenpunkten plus der dazu korrespondierenden synthetischen Referenzdaten eine Ellipse anzulegen, um aus deren LISSAJOUS-Ellipsen-Geometrieparameter auf die Phasenverschiebung zu schließen.

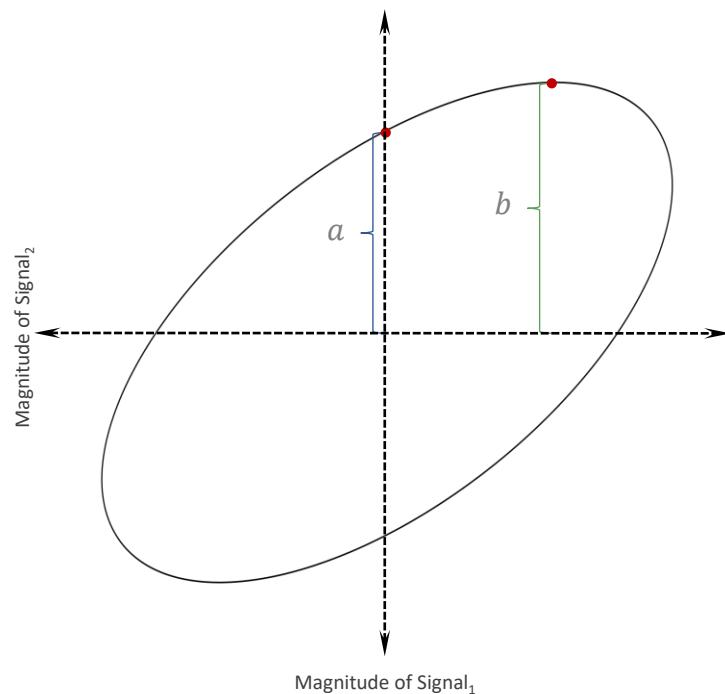


**Abbildung 13:** *Lissajous-Figuren.* A zeigt zwei phasenverschobene Sinus-artige Signale sowohl verrauscht, als auch clean. B zeigt die dazu korrespondierende Lissajous-Figur innerhalb eines Magnituden-Magnituden-Diagramms, sowohl hinsichtlich der verrauschten Signale durch die so entstandene Punktwolke, als auch der korrespondierenden unverrauschten Signale in Form der stetigen Ellipse.



# Methode III – LISSAJOUS-Ansatz

Um nun aus der angepassten Ellipse die Phasenverschiebung zu errechnen braucht es die Y-Werte bezüglich des Schnittpunkts der Ellipse mit der Y-Achse, sowie dem Maximalsten Y-Wert der Ellipse. Abb. 14 soll das etwas besser darstellen. Die beiden Parameter  $a$  und  $b$  stellen die beiden gesuchten Y-Werte dar. Damit ist es natürlich noch nicht getan. Denn um aus ihnen die Phasenverschiebung zu erhalten, muss noch etwas Trigonometrie betrieben werden.



Mit folgender trigonometrischer Gleichung lässt sich der Phasenwinkel zwischen den beiden Signalen errechnen:

$$\Delta\varphi = \arcsin\left(\frac{a}{b}\right) \quad (2)$$

Um nun aber den konkreten Zeitabstand in Tagen zu haben, muss der Phasenwinkel noch umgerechnet werden. Es gilt:

$$\Delta\varphi = 2 \cdot \pi \cdot f \cdot \Delta t \quad (3)$$

Da  $\Delta t$  unsere Phasenverschiebung als Zeitabstand ist, brauchen wir also nur die Gleichung umstellen und den Phasenwinkel durch den Arkussinus substituieren. Der Parameter  $f$  ist die Frequenz, welche wir durch  $\frac{1}{T}$  ersetzen können.  $T$  ist die Periodenlänge, welche ja in unserem Fall 365 Tage entspricht.

Der Faktor  $2 \cdot \pi \cdot f$  bzw.  $2 \cdot \pi \cdot \frac{1}{T}$  ist also eine (uns bekannte) Konstante,

$$\Delta t = \frac{T}{2 \cdot \pi} \cdot \arcsin\left(\frac{a}{b}\right) \quad (4)$$

**Abbildung 14:** Berechnung der Phasenverschiebung anhand der LISSAJOUS-Ellipse und ihrer Geometrieparameter  $a$  und  $b$ .

Der Parameter  $a$  ist der Schnittpunkt der Ellipse mit der Y-Achse. Der Parameter  $b$  ist der Maximalwert der Ellipse.



# Methode III – LISSAJOUS-Ansatz

Bevor wir aber die beiden LISSAJOUS-Ellipsen-Parameter  $a$  und  $b$  durch eine Regression schätzen bzw. vielmehr aus ihnen den Phasenverschiebungswinkel, reden wir noch etwas übers Filtern. Wir werden nämlich noch einmal den bereits schon einmal bewährten DBSCAN-Filter verwenden, um die Daten vorab etwas zu bereinigen.

Zudem werden wir auch diesmal nicht alle Datenpunkte benutzen, sondern eine zufällige Auswahl um so auf verschiedene Stichproben die Phasenverschiebung berechnen zu können.

In der nächsten Abbildung kann man das Ganze sehen. Es sind deutlich weniger Datenpunkte geplottet, da wir per Zufall einige Datenpunkte aus dem Gesamtumfang unserer Daten gezogen haben. Zudem sind einige Datenpunkte rot markiert. Diese wurden vom DBSCAN-Algorithmus als nicht valide erkannt und werden nicht weiter mit in die Analyse fallen.

Wer aufmerksam ist, dem wird aufgefallen sein, dass die Magnituden-Dimensionierung (Y-Achse) nicht mehr original sind.

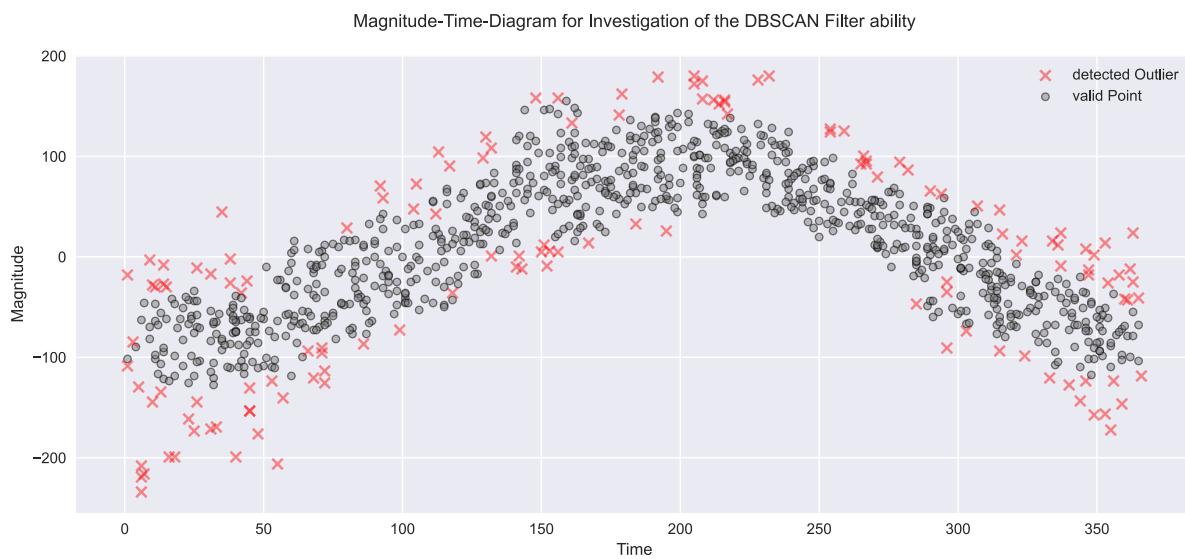


Abbildung 15: Zufällig gezogene 1000 Datenpunkte und DBSCAN-Ergebnisse.

Die rot markierten Punkte sind vom DBSCAN-Algorithmus erkannt wurden und werden aus der Stichprobe entfernt. Man kann sehr gut sehen, dass die meisten roten Punkte vom Muster abweichen und somit als Ausreißer gelten können. Auch in diesem Beispiel leistet der Filter gute Dienste.

Bevor wir die Daten zum DBSCAN schicken transformieren wir sie nämlich noch so, dass sie eine ähnliche Dimensionierung haben wie die der X-Achse.

Zudem ist die Amplitude nun so angeglichen, dass sie halbwegs gleichmäßig um  $y = 0$  schwingt.

Die Transformation soll zum einen dem DBSCAN-Algorithmus beim Finden der Ausreißer unterstützen, zum andern soll es aber für bessere Ergebnisse beim Schätzen der Phasenverschiebung sorgen.

Es ist für die Analyse nämlich theoretisch eigentlich wichtig, dass die Amplitude gleichmäßig um den Null-Wert seitens der Y-Achse schwingt. Denn das hat zur Folge, dass der Flächenschwerpunkt der LISSAJOUS-Ellipse nahe dem Koordinatenursprung liegen würde.

*Der Algorithmus zum Ausrechnen des Phasenwinkels, kann jedoch solche Verschiebungen hinsichtlich des LISSAJOUS-Ellipsen-Schwerpunkts über eine Transformation ausgleichen, sodass wir uns da dennoch keine Sorgen machen müssen, falls der Schwerpunkt nicht im Koordinatenursprung liegt.*

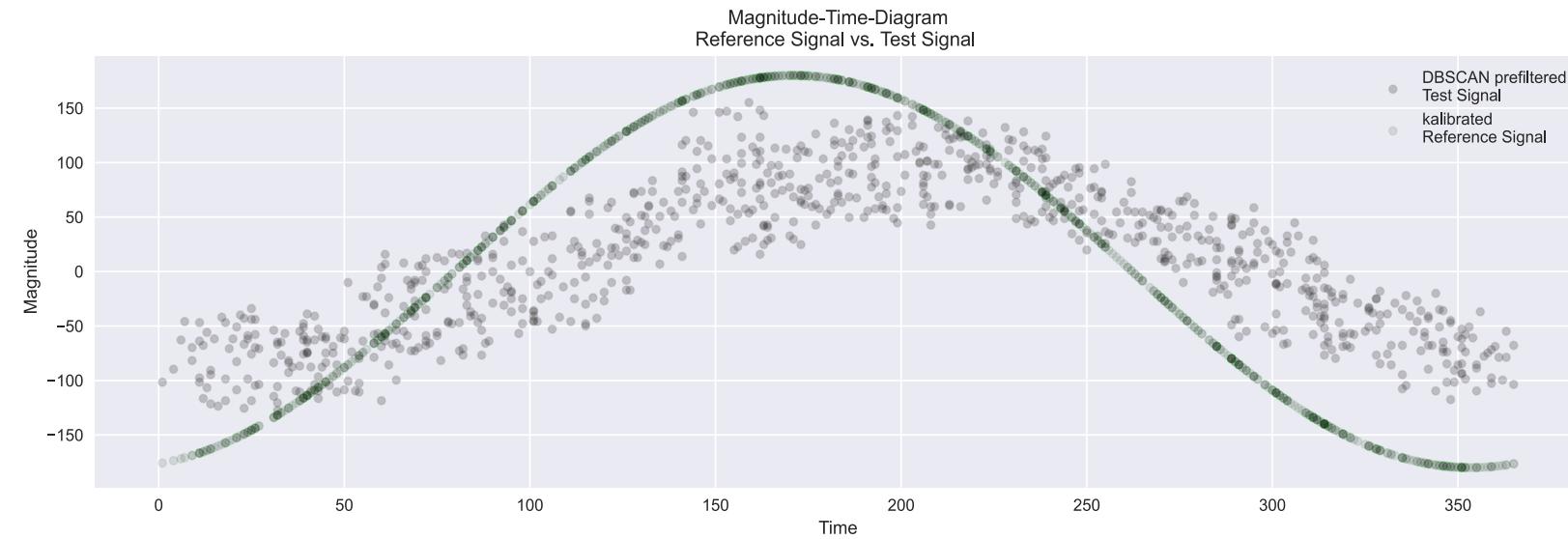
Falls Du Dich nun fragst warum wir einfach die Magnitudenwerte verändern dürfen. Wir untersuchen nur die Verschiebung der beiden Sinus-Funktionen. Phase und Magnitude sind aber immer unabhängig, weshalb wir die Y-Achse transformieren dürfen. Die X-Achse lassen wir jedoch unberührt.



# Methode III – LISSAJOUS-Ansatz

Nachdem die Stichprobe gefiltert wurde, schauen wir uns das Ergebnis noch zusammen mit den Datenpunkten der Referenzfunktion an. Die Referenzfunktion ist eine Sinusfunktion die wir bzgl. der Magnitude so angepasst haben, dass sie nahe an die Datenpunkten kommt. Das ist theoretisch gar nicht so wichtig, da wir ja bereits wissen, dass die Amplitudenstärke unerheblich ist. Allerdings bietet es sich dennoch an, dies zu tun, da es für den Algorithmus, welcher via Regression die Ellipsenfunktion anpasst, besser ist, mit ähnlich dimensionierten Werten zu rechnen um numerische Instabilitäten aus dem Weg zu gehen.

Zudem wurde der Phasenwinkel der Referenzfunktion so deklariert, dass das Maximum der Referenzfunktion beim längsten Tag liegt (Zeitindex = 173 → 21. Juni). Mit dieser Funktion können wir zu jedem Datenpunkt der täglichen mittleren Tagestemperatur nun einen Referenzamplitudenwert der täglichen Sonnenscheindauer zuordnen, um so mithilfe der Geometrieparamter aus der korrespondierenden LISSAJOUS-Ellipse auf die Phasendifferenz beider Signale schließen zu können.



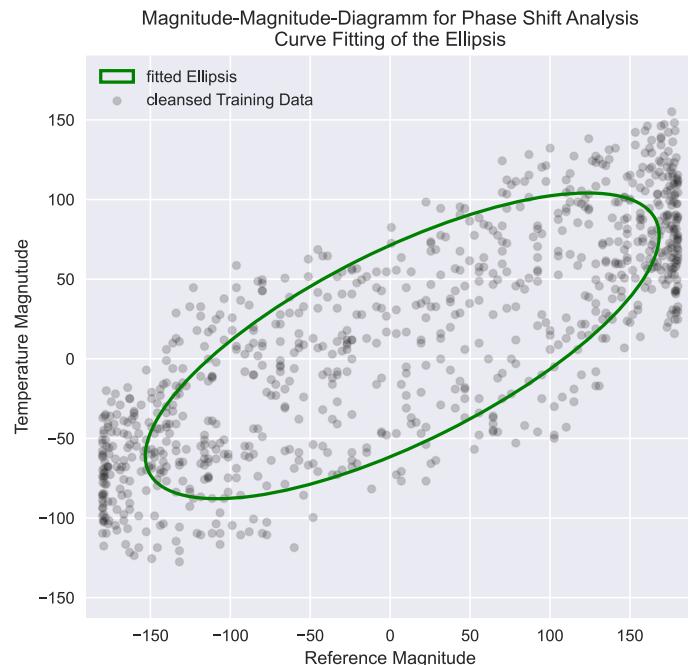
**Abbildung 16:** Zufällig gezogene und anschließend DBSCAN-gefilterte Datenpunkte und die korrespondierenden Punkte der Referenzfunktion als Ersatzfunktion für die tägliche Sonnenscheindauer.

Es kann anhand der grünen Datenpunkte der Referenzfunktion gesehen werden, dass es nicht konsistent für jeden Messtag einen oder mehrere Datenpunkt gibt, da wir nun nur noch zufällig gezogene Punkte betrachten. Wenn es keinen Datenpunkt in für einen bestimmten Zeitpunkt gab, so darf es natürlich auch keinen für die Referenzfunktion geben. Die Zeit spielt bei diesem Model nur eine implizite Rolle.

Die Amplituden sind jeweils unterschiedlich stark ausgeprägt, was jedoch keine Rolle spielt. Beide schwingen aber durch die Transformation der Datenpunkte annähernd gleichmäßig um den O-Punkt der Y-Achse. Dies ist wichtig, damit der Flächenschwerpunkt der LISSAJOUS-Ellipse möglichst nahe dem Koordinatenursprung liegt.



# Methode III – LISSAJOUS-Ansatz



**Abbildung 17:** Die Magnituden von Testfunktion und Referenzsinus als Scatterplot zusammen mit der darauf angepassten Lissajous-Ellipse.

Die Datenpunkte zeigen kein annähernd ideales Bild einer Ellipse. Zudem liegt der Flächenschwerpunkt der Punktfolge und auch der der angepassten Ellipse nicht im Koordinatenursprung.

Der Curve-Fitting-Algorithmus von BEN HAMMEL & NICK SULLIVAN-MOLINA schätzt dennoch eine Ellipse, welche gut in der Punktfolge zu liegen scheint.

Nun schauen wir uns noch ein exemplarisches Ergebniss einer angepassten Lissajous-Ellipse gefüttet an eine zufällige Auswahl unserer Datenpunkte an (siehe Abbildung 16). Die Magnituden sind als graue Punkte im Diagramm zu sehen. Trotz Filterung gibt es Anhäufungen im Bereich der Magnitudenmaxima. Man könnte dem natürlich durch eine weitere DBSCAN-Filterung entgegnen.

Nun, zugegeben, ein wirklich „ideales“ Bild einer Ellipse zeichnet sich bei unseren Datenpunkten nicht wirklich ab.

Wir machen trotzdem weiter und benutzen das Ellipsen-Anpassungs-Programm von HAMMEL & SULLIVAN-MOLINA und sehen, dass es dennoch eine Ellipse an die Punktfolge angleicht, welche relativ gut darin platziert wirkt.

Wie bereits erwähnt, interessieren uns zwar eigentlich die Geometrieparameter  $a$  und  $b$ . Aber nur deshalb, da sie implizit die gesuchte Information des Phasenverschiebungswinkels enthalten. Diesen müssten wir jetzt noch aus ihnen errechnen. Glücklicherweise ist das Programm aber um eine Implementierung dieses Schritts von den Autoren erweitert wurden.

Zudem müsste normalerweise der Flächenschwerpunkt der Ellipse im Koordinatenursprung liegen. In unserem Fall tut er das nach Abbildung 17 nicht. Allerdings funktioniert der Algorithmus von HAMMEL & SULLIVAN-MOLINA dennoch, da die Ellipse implizit vor dem Berechnen des Phasenwinkels bzgl. ihres Flächenschwerpunkts in den Koordinatenursprung verschoben wird.

Das Ganze wurde nun wieder nach dem schon aus der vorhergehenden Methode bekannten Monte-Carlo-Prinzip mehrmals auf einen zufällig gezogenen Teil unserer Datenpunkte gemacht. Das bedeutet, wir haben nun mehrere Modellwerte bzgl. der Verschiebungszeit und können daraus wieder einen Mittelwert des Modells und eine Varianz bzw. Standardabweichung des Modells berechnen um eine Einschätzung bzgl. der Modellunsicherheit machen zu können.



# Methode III – LISSAJOUS-Ansatz

Bei dem ganzen Aufwand scheint dieser banale Boxplot (Abb. 18), welcher unsere Modelergebnisse statistisch repräsentiert fast schon etwas kläglich. Aber dafür sind die Ergebnisse nicht die Schlechtesten.

Es konnten für die Analyse wieder 18263 Datenpunkte genutzt werden. Insgesamt wurde das Modell 100 mal durchlaufen wobei immer 1000 Punkte vorab zufällig ohne zurücklegen beprobt wurde. Allerdings darf dabei nicht vergessen werden, dass einige von ihnen nicht den Weg durch den DBSCAN-Filter geschafft haben.

Auf der Y-Achse ist der Phasenverschiebungswert [Tage] aufgetragen. Der mittlere Phasenverschiebungswert beläuft sich also auf ca. 22 Tage (Median = 22, Mittelwert = 22.26). Die Standardabweichung wurde mit 1,7 Tagen beziffert. Das die Streuung so gering ist, ist schonmal ein sehr gutes Zeichen.

Allerdings ist das nur ein relativer Wert, nämlich in Form der Phasenverschiebung zum Referenzwert. D.h. wir müssen nun noch den Absolutwert daraus errechnen, welcher sich also aus dem Tag-Index des 21. Juni plus dem mittleren Modellwert ergibt.

Als Referenzwert gilt der 21. Juni. D.h. der so errechnete Tag bzgl. der mittleren täglichen Durchschnittstemperatur im jährlichen Maximum würde laut diesem Modell und der errechneten Phasenverschiebung also auf den 13. Juli fallen. (angepasster Mittelwert  $\bar{x}_L = 195.26$ )

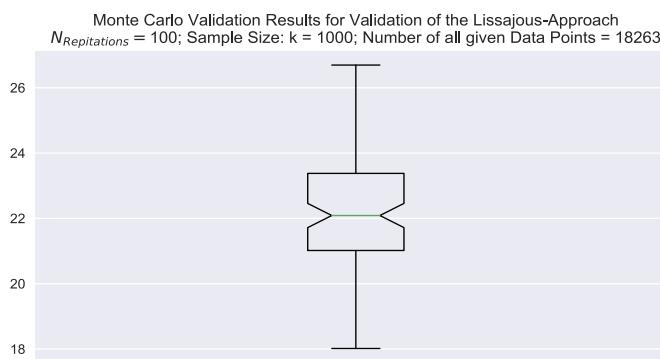


Abbildung 18: Ergebnisse aller Monte-Carlo-Durchläufe bzgl. des Lissajous-Models.

Die mittlere Phasenverschiebung liegt gemessen am Median bei 22 Tagen, woraus sich das Datum der maximalsten Jahrestemperatur zum 13. Juli errechnen lässt.



# Alle Modelle und Ergebnisse im Vergleich

Da wir nun alle 3 Modelle getestet und validiert haben, geht es nun ans Küren des Gewinners.

Allerdings wird das nicht ganz so leicht. Denn anders als bei dem Tag der maximalen Sonnenscheindauer, bei dem wir einfach den (bekannten) Tag der astronomisch längsten Sonnenscheindauer als Referenzwert/Erwartungswert genommen haben, gibt es für den Tag der maximalsten Jahrestemperatur scheinbar keinen solchen Referenzwert in der Literatur. Jedenfalls habe ich keinen in irgendwelchen Quellen finden können.

Auf Wikipedia stieß ich aber zumindest auf eine Tabelle mit einigen gemessenen heißesten Tagen in Deutschland (siehe Abb. 19). Wir werden diese Tage als Referenz benutzen um zumindest etwas zum Vergleichen zu haben. *Und ich nehme es schon vorweg, das Ganze wird statistisch sehr grenzwertig!* Wir können damit allerhöchstens schauen, ob unsere Modelle halbwegs plausibel sind, ohne aber dabei wirklich eine große Sicherheit zu haben. Also nagelt mich nachher bitte nicht daran fest!...

State	Extreme Maximum			Tag Index
	Temperature	Location	Date	
Baden-Württemberg	40.2 °C (104.4 °F)	Bad Mergentheim <sup>[2]</sup>	August 7, 2015	219
Bavaria	40.3 °C (104.5 °F)	Kitzingen <sup>[4]</sup>	August 7, 2015	219
Berlin	38.6 °C (101.5 °F)	Tegel Airport <sup>[7]</sup>	June 30, 2019	181
Brandenburg	39.2 °C (102.6 °F)	Lübben <sup>[9]</sup>	August 9, 1992	222
Bremen	37.6 °C (99.7 °F)	Bremen Airport <sup>[11]</sup>	August 9, 1992	222
Hamburg	37.8 °C (100.0 °F)	Wandsbek <sup>[13]</sup>	July 20, 2006	201
Hesse	40.2 °C (104.4 °F)	Westend, Frankfurt <sup>[15]</sup>	July 25, 2019	206
Lower Saxony	39.7 °C (103.5 °F)	Meppen <sup>[17]</sup>	July 25, 2019	206
Mecklenburg-Vorpommern	38.0 °C (100.4 °F)	Ueckerbrücke <sup>[19]</sup>	August 1, 1994	213
North Rhine-Westphalia	41.2 °C (106.2 °F)	Duisburg & Tönisvorst <sup>[21]</sup>	July 25, 2019	206
Rhineland-Palatinate	40.4 °C (104.7 °F)	Andernach & Bad Neuenahr-Ahrweiler <sup>[23]</sup>	July 25, 2019	206
Saarland	40.3 °C (104.5 °F)	Nennig <sup>[24]</sup> Perl	August 8, 2003	220
Saxony	39.8 °C (103.6 °F)	Hosterwitz, <sup>[26]</sup> Dresden	August 20, 2012	233
Saxony-Anhalt	38.8 °C (101.8 °F)	Gardelegen <sup>[28]</sup>	July 4, 2015	185
Schleswig-Holstein	37.2 °C (99.0 °F)	Grambek <sup>[29]</sup> near Mölln	July 4, 2015	185
Thuringia	38.8 °C (101.8 °F)	Jena <sup>[32]</sup> Astronomical Observatory	June 30, 2019	181

Abbildung 19: Einige gemessene Maximaltemperaturtage in Deutschland mit den korrespondierenden Tages Indizes als Erweiterung der Originaltabelle (verändert durch den Autor um die zusätzliche grün markierten Spalte „Tag Index“).

Quelle: [https://de.wikipedia.org/wiki/Liste\\_der\\_Temperaturrekorde\\_in\\_Deutschland](https://de.wikipedia.org/wiki/Liste_der_Temperaturrekorde_in_Deutschland) unter der Üblichen CC-Lizenz.

Wir werden die Daten jeweils auf ihren Tagesindex umrechnen (siehe zusätzliche Spalte *Tag Index*) und daraus den Mittelwert aller bilden. Diesen Mittelwert benutzen wir dann als Referenzwert. Die rot markierten Zahlen beziehen wir dabei nicht mit in die Schätzung ein, da dies Duplikate sind. Es wurden also zum gleichen Datum die Maximaltemperatur an verschiedenen Orten in Deutschland gemessen. Es würde also keinen Sinn machen, diese Werte mehrfach einzubeziehen und die (ohnehin recht seichte) Statistik sogar noch zu verzerren.

Für unseren Referenzwert, den wir nun  $\hat{\mu}_{ref}$  nennen, ergibt sich:  $\hat{\mu}_{ref} = 208.9$ . Diesen verwenden wir nun, um neben dem relativen Vergleich der Modellergebnisse untereinander auch einen absoluten Vergleich zu haben. Auch wenn er nicht ideal ist! Er dient uns also als Ersatz für den Erwartungswert. Denn unser Stichprobenumfang ist mit  $n = 9$  wirklich lächerlich klein, und lässt somit keine wirklich sichere Schätzung zu. Zum Anderen benutzen wir hier zudem keine Tagesdurchschnittstemperaturen, sondern Tagesmaximumtemperaturen. Wir wollen aber den Tag der maximalen Tagesdurchschnittstemperatur mit unseren Modellen Schätzen. Es ist zwar intuitiv geschätzt wahrscheinlich, dass der Tag mit der jährlichen maximalsten Tagesdurchschnittstemperatur nahe dem Mittewert aller gemessenen Maximaltemperaturen liegt, aber wir werden im letzten Kapitel seitens der Ergebnisinterpretation noch über Tagestemperaturen diskutieren.

Halten wir also fest, dass dieser Wert hier einfach nur nochmal eine Tendenz bzgl. der Modellplausibilität zeigen soll. Letztlich ist es eigentlich auch gar nicht nötig, da wir innerhalb unserer Modellvalidierungen basierend auf sehr hohen Stichprobenumfängen schon eine sehr hohe Sicherheit haben.

Ich mag es allerdings, Modelle immer noch zusätzlich durch unabhängige Literaturwerte zu vergleichen.

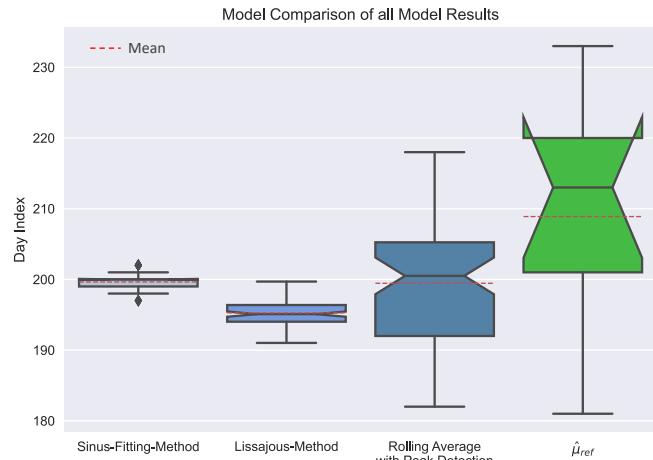


# Alle Modelle und Ergebnisse im Vergleich

Kommen wir nun also zum krönenden Abschluss des Ganzen und schauen uns alle Ergebnisse nochmal im direkten Vergleich miteinander sowie mit unserem Referenzwert an.

Abbildung 19 ist ein Boxplot aller Modellergebnisse bzgl. ihrer Mittelwerte, Mediane und Streuung. Es ist zu sehen, dass alle Modelle hinsichtlich ihrer mittleren Ergebnisse schonmal relativ zueinander sehr nahe liegen. Insbesondere das Modell der angepassten Sinusfunktion und das des gleitenden Mittelwerts teilen sich beide fast den gleichen Mittelwert (rote gestrichelte Linien). Zudem liegen sie auch nahe bei unseren Referenzwert  $\hat{\mu}_{ref}$ , den wir (vorsichtig) als Erwartungswert nutzen, repräsentiert durch die letzte grüne Box.

Neben dem Box-Plot haben wir noch die untere Tabelle (Tab. 2) mit den Mittelwerten aller Modelle, dem Referenzwert sowie den absoluten Abweichungen jeweils der Mittelwerte zum Referenzwert um sie somit auch numerisch untereinander vergleichen zu können.



**Abbildung 19:** Statistischer Vergleich aller 3 Modelle.

Die grüne Box repräsentiert die Schätzung des Ersatzwerts. Die roten Linien sind die Mittelwerte, während die Einschnürungen die Mediane repräsentieren.

$\hat{\mu}_{ref}$	$\bar{x}_{sin}$	$ \hat{\mu}_{ref} - \bar{x}_{sin} $	$\bar{x}_L$	$ \hat{\mu}_{ref} - \bar{x}_L $	$\bar{x}_R$	$ \hat{\mu}_{ref} - \bar{x}_R $
208.9	199.6	9.3	195.3	13.6	199.5	9.4

**Tabelle 2:** Statistischer Vergleich aller 3 Modelle durch die Mittelwerte in absoluter Differenz zum Referenzwert als Ersatz für den Erwartungswert.

$\bar{x}_{sin}$  := Mittelwert der Modellergebnisse basierend auf der Sinusfunktionsanpassung,  $\bar{x}_L$  := Mittelwert der Modellergebnisse basierend auf dem LISSAJOUS-Ansatz,  $\bar{x}_R$  := Mittelwert der Modellergebnisse basierend auf den Ansatz des gleitenden Mittelwerts mit anschließender Peak Detection. Die absoluten Differenzen sind durch die Betragsschritte gekennzeichnet.

Im relativen Vergleich der Modelle untereinander weichen die beiden Mittelwerte der Modelle: angepasste Sinusfunktion und gleitender Mittelwert mit anschließender Maxima-Detektion nur um 2.4 Stunden voneinander ab. Das ist enorm gut und spricht aufgrund der sehr hohen Stichprobenumfänge für eine sichere Validität beider. Das Modell mit der angepassten Sinusfunktion weist jedoch im Vergleich zu dem Anderen nur eine geringere Varianz um den Mittelwert auf. Dabei weicht das LISSAJOUS-Modell um ca. 4 Tage von beiden ab.

Nun noch zum weniger idealen Absolutvergleich, bei dem wir unseren Referenzwert  $\hat{\mu}_{ref}$  als Vergleichsmetrik nehmen.

Die Abweichungen der Mittelwerte zum Referenzwert  $\hat{\mu}_{ref}$  sind insgesamt relativ groß liegen aber alle im Streubereich des Referenzwerts. Die größte Abweichung besitzt das Modell basierend auf der LISSAJOUS-Methode mit rund 14 Tagen. Die beiden anderen Modelle weichen bzgl. ihrer Mittelwerte nahezu gleich mit rund 10 Tagen vom Referenzwertmittel ab. Dieser Vergleich soll aber auch nur zeigen, ob wir tendenziell mit den Modellen in einem sinnvollen Bereich liegen oder nicht völlig daneben sind. Allein die sehr große Varianz der zur Schätzung des Referenzwerts zugezogenen Werte zeigt, dass hier viel Vorsicht geboten ist. Zudem war der Stichprobenumfang viel zu gering als das der Mittelwert die Grundgesamtheit sicher repräsentieren kann.

Eine Plausibilität aller Modelle ist jedoch definitiv schon durch den relativen Vergleich der Modelle untereinander gegeben!



# Hypothesentest

...und? Stimmt denn die Hypothese nun?

Das Schlechteste Modell bemessen an der Standardabweichung ist das des gleitenden Mittelwerts mit anschließenden Maxima-Detektion. Alle anderen brauchen wir deshalb schon gar nicht weiter betrachten. Wir können nämlich behaupten, dass unter den gegebenen Umständen jedes bessere Modelle erst recht dann funktioniert, wenn eben auch das schlechteste die Hypothese bestätigt. Klingt logisch, oder? Diese Logik ist allerdings dem Fakt geschuldet, dass die beiden anderen Modelle hinsichtlich ihrer Prognose plus Unsicherheit innerhalb der Unsicherheit von diesem Modell liegen.

Zum beweisen unserer Hypothese formulieren wir nun erstmal eine sogenannte Null-Hypothese  $H_0$ , welche unsere eigentliche Hypothese (mit großer Sicherheit) widerlegen würde, sofern  $H_0$  wahr wäre. Es könnte nämlich einfach Jemand kommen und clever behaupten, dass die Modellunsicherheit bemessen an dessen Standardabweichung einen Zeitbereich definiert in dem der Tag der tatsächlichen maximalsten Sonnenscheindauer also der 21. Juni rein zufällig liegt und der modellierte Tag somit gar kein signifikanter Zeitpunkt wäre. Als Unsicherheitszeitraum schlägt er nun das Intervall  $[\mu - 1.65 \cdot \sigma_{\bar{x}_R}; \mu]$  vor, wobei  $\mu$  der bekannte Erwartungswert abbildend auf den 21. Juni ist und  $\sigma_{\bar{x}_R}$  die Standardabweichung aller Mittelwerte unserer Monte-Carlo-Modell-Simulationen. Denn dann, mit einer Wahrscheinlichkeit von 95 %, ist  $H_0$  nämlich genau dann wahr, wenn nun also der 21. Juni in diesem Unsicherheitszeitraum liegen würde. Und das führte zum Umkehrschluss, dass wir mit unserer Hypothese also (sehr wahrscheinlich) falsch liegen...und viel Arbeit -nicht umsonst!<sup>1</sup>- aber mit nur sehr ernüchternden Ergebnis gemacht hätten.

Aber rechnen wir mal wie es denn nun ist...

Der Tag der maximalsten Durchschnittstemperatur den das Model voraussagt ist ja der 17. Juli. Da die 1.65 -fache Standardabweichung aller Modellmittelwerte 14.30 Tage beträgt, datiert sich dieser Fall auf den rund 2. Juli, weil der 17. Juli minus der rund 15 Tage bzgl. der 1.65 -fachen Standardabweichung den 2. Juli ergeben wir haben hier sogar großzügig zu unserer (hypothetischen) Ungunst aufgerundet, **weshalb die 95% Sicherheit sogar noch etwas erweitert werden**. Der Unsicherheitsbereich ist demnach also der 2. Juli bis 17. Juli.

Da aber der Tag mit der längsten Sonnenscheindauer der 21. Juni ist und somit nicht im Bereich zwischen dem 2. Juli und dem 17. Juli liegt, können wir  $H_0$  getrost ablehnen und uns sehr sicher sein, dass unsere Hypothese wahr ist.

Das hatten wir uns natürlich schon vorher gedacht, da es schon sehr offensichtlich ist, wenn man sich die Daten anschaut. Aber wenn man Wissenschaft vernünftig betreibt, muss man Hypothesen auch „astrein“ beweisen.

*Tatsächlich stehe ich Hypothesentests aber seeehr kritisch entgegen. Macht euch mal den Spaß, und versucht die Ergebnisse publizierter Experimente etwa aus dem Bereich der Psychologie und Sozialwissenschaften<sup>2</sup> angegeben mit angeblich  $p < 0.05$  zu reproduzieren. Laut „Hypothesentest-Theorie“ sollte euch das also mit einer Wahrscheinlichkeit  $\geq 95\%$  gelingen. Aber bei so einigen wird es nicht der Fall sein! Guck mal dazu einfach [hier](#)<sup>3</sup> rein. Wenn der Stichprobenumfang nur groß genug ist, kann ich Dir die aberwitzigsten Hypothesen durch Stichprobenbeschneidung mit  $p < 0.05$  „beweisen“! Alles nur eine Frage des Samplings!*

*Leider ist der p-Wert zu einer Metrik verkommen welche zusammen mit einer weiteren – aus meiner Sicht – sehr fragwürdigen Metrik dem Hirsch-Index, aus Wissenschaft immer mehr etwas macht, was mit Wissenschaft nur noch wenig gemein hat. Ein ALEXANDER VON HUMBOLDT dreht sich wohl fast ununterbrochen im Grab um...*

1: Eine Falsifizierung, also die Wiederlegung einer Hypothese ist in der Wissenschaft - aus meiner Sicht – genauso wichtig und interessant wie deren Kompliment. Nur leider werden die wenigsten wissenschaftlichen Journals (vermutlich gar keins) einen Artikel veröffentlichen, indem Du Deine Hypothese nicht beweisen konntest. Zudem wirft es ja auch ein „schlechtes Licht“ auf Dich, weil Deine Vermutung falsch war... Aber eigentlich ist es doch schade, bedenkt man, dass es vielleicht anderen WissenschaftlerInnen Arbeit ersparen würde, wenn sie die gleiche Idee hätten und nun nichtsahnend auf die gleichen Ergebnisse stießen... Tja, in unserer perfekten Welt muss eben immer alles perfekt sein...

2: Sorry an alle Psychologinnen und SozialwissenschaftlerInnen welchen diese Jacke nicht passt! Aber das Verhältnis von den Schwarzen Schafen in Sachen „p-Wert-Gauñereien“ zur Gesamtmenge ist in eurer „Zunft“ leider bekannter Maßen besonders hoch. Aber unter dem derzeitigen Stand der wissenschaftlichen Kultur (Publikationsdruck, befristete Arbeitsverhältnisse etc.) auch irgendwie verständlich... siehe dazu auch Fußnote 1.

3: <https://www.uzh.ch/blog/hbz/2019/11/19/p-hacking-kein-kavaliersdelikt/>



# Lessons Learned

Alle Wege führen nach Rom ... aber nicht jeder ist der Beste!

Wir konnten 3 Verschiedene Modelle evaluieren und statistisch validieren. Dabei konnten wir feststellen, dass es nur geringfügige Unterschiede in ihrer Güte gibt. Zugegeben, die Vergleichsstatistik mit dem Ersatzwert  $\mu_{ref}$  als Erwartungswert ist etwas gewagt. Es schwingt eine kleine Unsicherheit bei der ganzen Sache mit. Bedenkt man aber, dass alle 3 Modell sehr ähnliche Ergebnisse liefern, welche zudem auch noch sehr nah an dem Referenzwert liegen, sollten wir uns dennoch ziemlich sicher sein dürfen.

Die LISSAJOUS-Methode war wohl etwas zu kreativ. Zumaldest aber gibt es hinsichtlich der Streuung keine sehr große Unsicherheit. Aber die Taktik der Transformation in einen anderen mathematischen Raum um bessere Ergebnisse zu erhalten, scheint nicht ganz aufgegangen zu sein. Das Modell ist relativ schwierig und komplex. Das etwas schlechte Ergebnis steht hier nicht im optimalen Verhältnis zur Modellkomplexität.

Das einfache Modell auf Basis des optimierten gleitenden Mittelwerts mit anschließender Maximum-Erkennung bietet die größte Unsicherheit obwohl der Mittelwert am nahesten am Erwartungswert liegt. Das ist ziemlich interessant, da wir hier zwei statistische Metriken haben, die bei der Modellgüteentscheidungen quasi gegeneinander arbeiten. Das Modell trifft den Erwartungswert - wenn er den überhaupt richtig ist - im Mittel am nahesten aber dennoch ist die Unsicherheit weitaus größer, als bei den anderen Modellen.

Die Anpassung einer Sinusfunktion an die Beobachtungsdaten bietet ein ziemlich gutes Modell. Eine sehr geringe Streuung ist schon ein sehr gutes Zeichen, dass dieses Modell gut performt. Aber beim Testen der Methode auf die Sonnenscheindauer wurde der (sichere) Erwartungswert um nur einen Tag verfehlt.

Also: Das Modell rockt !!!

Dennoch mag ich den LISSAJOUS-Ansatz (auf recht subjektive Weise) am liebsten. Es war ein Versuch wert ;)...und unter totales Scheitern kann er definitiv nicht verbucht werden.

Die Hypothese konnten wir ziemlich eindeutig beweisen!

...bleibt nun aber noch offen, warum dieser zeitliche Versatz beider Wetterparameter eigentlich existiert, oder?



# Ergebnisinterpretation

Warum die zeitliche Verschiebung?

Nun sind wir den gesamten Werdegang einer Datenanalysearbeit durchgegangen.

Wir konnten die Daten aufbereiten und haben sie visuell unter die Lupe genommen.

Dann haben wir eine Hypothese formuliert.

Anschließend haben wir die Daten anhand verschiedener Methoden analysiert und Modelle designt, die wir auch statistisch vernünftig validiert haben.

Die Hypothese, konnten wir mit ziemlicher Sicherheit durch Formulierung und sicherem Verwurf einer Nullhypothese klar beweisen.

Job erfüllt !!!

...aber so richtige befriedigend ist es eigentlich doch noch nicht, wenn wir nicht auch wissen, warum denn nun diese zeitliche Verschiebung existiert, oder?

Begeben wir uns nun also in den letzten Teil unserer kleinen „Reise“, wobei wir die abgeschlossene Datenanalyse hinter uns lassen und nun den Hintergrund der bewiesenen Hypothese bzgl. unserer Daten noch etwas beleuchten.



# Ergebnisinterpretation

Warum die zeitliche Verschiebung?

Wir konnten relativ eindeutig nachweisen - egal welches Modell wir nun nutzen, dass es definitiv einen zeitlichen Versatz zwischen der täglichen mittleren Temperatur und der Sonnenscheindauer bzgl. ihrer Maximalwerte gibt. Zudem konnten wir zeigen, dass es signifikante Zeitpunkte also bestimmte Tage bezogen auf die jährlichen Maxima der beiden Wetterparameter gibt und sich beide Wetterparameter durch ein Sinusfunktion abbilden lassen.

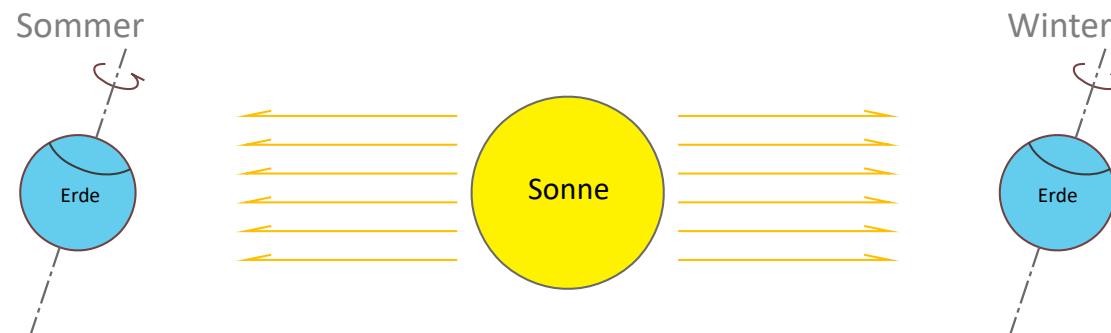
Zwar liegen die Maxima schon relativ dicht beieinander, aber die maximale mittlere Tagestemperatur „läuft“ um ca. einen Monat hinterher.

Der Grund dafür ist nicht nur die Sonnenscheindauer als solches. Schließlich scheint in den Polarsommern/wintern oberhalb der Polargrenzen die Sonne fast Rund um die Uhr. Jenseits dieser Grenzen ist es aber dennoch bitterkalt, oder? Die Sonnenscheindauer ist nämlich nicht der einzige Faktor. Die Erde ist bekanntlich bezüglich ihrer Rotationsachse relativ zur Umdrehungsebene zwischen Sonne und Erde um etwa  $23,5^\circ$  geneigt. Das führt dazu, dass im Sommer die Strahlungsichte größer ist als im Winter.

Was führ ein Ding?

Stell es Dir folgend vor: Wenn die (annähernd parallelen) Sonnenstrahlen auf die Erde treffen, dann fallen somit auf eine Test-Fläche (kleiner beliebiger Ausschnitt) der Erdoberfläche eine bestimmte Anzahl von Photonen<sup>1</sup> (Lichtwellen/Lichtteilchen...ja was denn nun?) pro Zeiteinheit. Während nun aber die Anzahl dieser Photonen, welche pro Zeiteinheit auf die gesamte zugewandte Erdoberfläche treffen halbwegs konstant ist, ebenso der Betrag der Testfläche, ist aber die Anzahl der Photonen die in gleicher Zeit auf die Testfläche fallen abhängig vom Breitengrad und der Position der Erde zur Sonne...und eigentlich auch noch Bedeckung des Himmels, atmosphärische Effekte, Sonnenaktivität etc., aber die genannten vernachlässigen wir jetzt mal großzügig.

Ich denke, es ist jetzt immer noch nicht ganz klar was ich meine, oder? Deswegen hier mal ein kleiner Versuch die Erde und Sonne als kinematisches Modell darzustellen (siehe Abbildung 20). Wichtig ist jetzt zu erkennen, dass die Rotationsachse abhängig von Sommer- und Winterzeit anders zur Sonne ausgerichtet ist. Denn der Rest folgt auf der nächsten Seite...



**Abbildung 20: Erde-Sonne-Beziehung hinsichtlich der Rotationsachse im Winter und im Sommer.**  
Die Erde ist bezogen auf ihre nördliche Halbkugel im Sommer der Sonne etwas zugewandter als im Winter.  
Für die Südhalbkugel gilt dann genau das Gegenteil, weshalb es dort im Sommer kälter ist

1: <https://de.wikipedia.org/wiki/Photon>



# Ergebnisinterpretation

Warum die zeitliche Verschiebung?

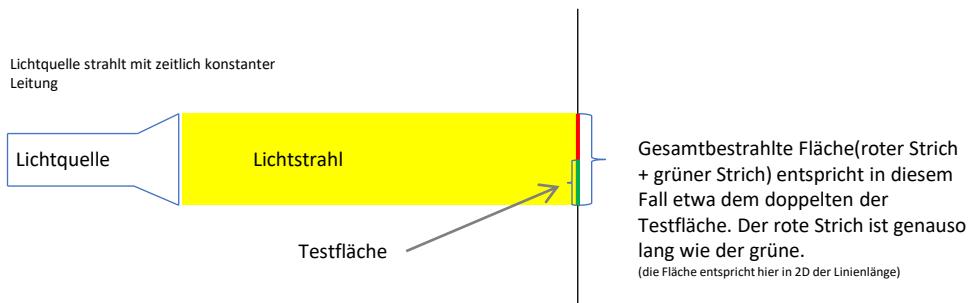
Im Sommer ist die Nordhalbkugel bedingt durch den Neigungswinkel der Rotationsachse der Sonne etwas zugeneigt, während sie im Winter der Sonne im gleichen Maße abgeneigt ist.

Wenn nun die Sonnenstrahlen auf die Erde gesendet werden, verteilen sie sich im Winter (auf der Nordhalbkugel) auf mehr Fläche als im Sommer.

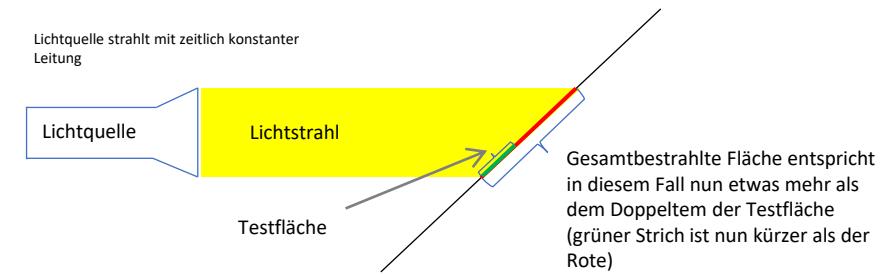
Stell es Dir gedanklich wie eine Taschenlampe vor, die Du auf ein großen Blatt Papier richtest. Wenn das Blatt normal zur Strahlenachse ausgerichtet ist, Du also die Taschenlampe direkt drauf hältst, dann sollte der auftreffende Lichtstrahl auf dem Papier gedacht einen Kreis abbilden. Wenn Du nun aber das Blatt relativ zur Strahlenachse neigst, zeichnet sich eine Ellipse auf dem Blatt ab, welche verglichen zum Kreis eine größere Fläche hat. Es treffen alle Strahlen zur gleichen Zeit nun also auf eine größere Gesamtfläche auf.

Zum noch besseren Verständnis malst Du jetzt gedanklich eine kleine Testfläche auf das Blatt. Das kann also etwa ein kleiner Kreis oder was auch immer sein. Im ersten Szenario leuchtest Du wieder direkt also „gerade“ auf das Blatt Papier, und die Strahlen treffen direkt drauf, wobei auch Deine Testfläche mitbestrahlt wird.

Nun das Ganze wieder mit gedacht geneigtem Blatt. Deine Testfläche wird sich um ihren Betrag natürlich nicht geändert haben nur weil Du mit der Taschenlampe nun schräg auf das Blatt leuchtest. Aber es treffen zu jedem Zeitpunkt nun weniger Strahlen auf Deine Testfläche ein, da sie sich die Strahlen ja auf mehr Gesamtfläche verteilen (müssen).



**Abbildung 21:** Taschenlampenstrahl normal auf ein Blatt gerichtet – Modellhaft (Papierebene relativ zu Taschenlampe nicht geneigt). Die Fläche der auftreffenden Strahlen wird repräsentiert durch die Länge der roten Linie. Die Grüne Linie repräsentiert eine Testfläche welche in jedem Szenario immer den gleichen Betrag aufweist.



**Abbildung 22:** Taschenlampenstrahl geneigt auf ein Blatt gerichtet – Modellhaft (Papierebene relativ zur Taschenlampe geneigt). Der auftreffende Strahl bildet nun eine größere Fläche ab. Die Testfläche bekommt nun pro Zeiteinheit weniger Photonen ab, da diese sich auf eine größere Gesamtfläche verteilen.



# Ergebnisinterpretation

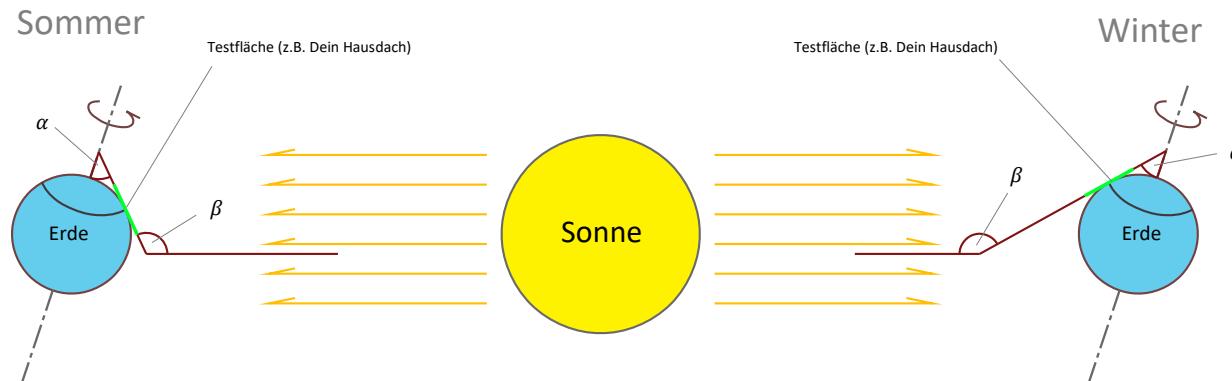
Warum die zeitliche Verschiebung?

Die Strahlungsdichte ist gegeben durch die Strahlungsleistung der Sonnenstrahlung pro Fläche. Wir haben von Photonen gesprochen. Stell sie Dir wie kleine „Wärme-Pakete“ vor, die kontinuierlich auf die Erde eintreffen. Da das „Erwärmungspotenzial“ des Bodens auch abhängig ist von der Strahlungsdichte, können wir nun schlussfolgern, dass sich der Boden dann besser erwärmt, wenn seine Fläche der Sonne so direkt wie möglich zugewandt ist. Genau deshalb ist es am Äquator nämlich auch gleichmäßig warm. Die effektive Bestrahlungsfläche hinsichtlich der Zeit verändert sich dort nämlich nur sehr geringfügig.

Würden wir die Erdneigung völlig außer Acht lassen bzw. gäbe es sie gar nicht, dann würde die effektive Bestrahlungsfläche zeitlich konstant sein und nur vom Breitengrad bzgl. der Veränderung abhängen. Wir erinnern uns an das Modell mit der Taschenlampe. Das hatte nur eben noch keine Krümmung der Fläche berücksichtigt. Die Krümmung der Erdoberfläche bewirkt nun also eine Vergrößerung der effektiv bestrahlten Fläche in Abhängigkeit des Breitengrads. D.h. je weiter man sich vom Äquator entfernt, desto weniger Strahlen treffen auf die Erdoberfläche in gleicher Zeiteinheit auf. Ergo erwärmt sich der Boden umso schlechter, je weiter wir uns dem Äquator entfernen. Und je näher wie die Polen kommen, umso kälter wird es eben. Und selbst dann, wenn die Sonne nahe dem Pol im z.B. Polarsommer auf der Nordhalbkugel über dem Polarkreis gar nicht richtig untergeht.

Vereinen wir jetzt unsere beiden Veranschaulichungen von Erde-Sonne-Kinematik aus Abbildung 19 mit dem [Taschenlampenmodell](#) (siehe Abbildung 21 und 22) und lassen aber die Sonne nun den Job der Taschenlampe übernehmen, dann sehen wir, dass eine konstant ausgerichtete Testfläche (grüner Strich) relativ zur Strahlenachse der Sonnenstrahlen im Sommer einen kleineren Winkel  $\beta$  aufweist, als im Winter. Die Testfläche könnte etwa das Dach Deines Hauses sein. Das verändert nämlich relativ zu irgendwelchen irdischen Bezügen nie seinen Winkel...also hoffentlich!!! Deswegen bleibt der Winkel  $\alpha$  zu jedem Zeitpunkt immer der gleiche. Das ist hier in unserem 2-D-Modell der Winkel  $\alpha$  zwischen der Rotationsachse und Deinem Hausdach. Dein Hausdach ist nun im Sommer der Sonne etwas zugeneigter und bekommt daher mehr Strahlen bzw. „Wärme-Pakete“ pro Zeiteinheit ab, als das im Winter der Fall ist.

Das ist im Übrigen auch der Grund, warum die Schatten im Winter länger sind, als im Sommer und es genauso wenig sinnvoll ist einem Inuit einen Kühlschrank zu verkaufen als auch Solarpanels für sein Dach ;).



**Abbildung 23:** Erde-Sonne-Konstellation bzgl. der Jahreszeiten und der Ausrichtung einer Testfläche.

Die Testfläche befindet sich in jedem Fall am gleichen Breitengrad unverändert und bleibt damit relativ zur Rotationsachse im gleichen Winkel  $\alpha$ . Jedoch ist der Winkel  $\beta$  zwischen Testflächenebenen und der Sonnenstrahlenausrichtung im Winter größer als im Sommer, also zeitabhängig. Dies führt dazu, dass im Sommer, pro Zeiteinheit mehr Sonnenstrahlen auf der Testfläche Landen, als im Winter.



# Ergebnisinterpretation

Warum die zeitliche Verschiebung?

Das ist der tatsächliche Grund, warum es im Winter bei uns auf der Nordhalbkugel kälter ist, als im Sommer. Lass Dir bitte keinen Bären aufbinden, es hätte was mit der Ellipsen-Bahnkurve der Erde zur Sonne zu tun. Diese existiert natürlich wirklich, aber da die Strahlen in guter Näherung parallel sind und im „fast“ Vakuum des Weltalls so gut wie gar nicht gedämpft werden, ist es zumindest in dieser Dimensionierung fast egal wie weit sich die Erde von der Sonne entfernt.

Tatsächlich spielt die Kugelgeometrie eines gleichmäßig strahlenden Objekts eine Rolle auf die Strahlungsdicht in Relation zum Abstand des strahlenden Objekt, da ja die Strahlen tatsächlich über die Distanz divergieren, also eigentlich nicht parallel verlaufen. Aber dieser Effekt ist bei dem Längenunterschied der beiden Radien der elliptischen Bahnkurven fast vernachlässigbar. Er spielt in der gesamten Wärmebilanz nämlich nur eine sehr geringe Rolle.

Zudem ist es so, dass die Radien bzgl. der Sommer-Winter-Zeit nicht fest stehen und der geringste Abstand (Perihel) zwischen Sonne und Erde derzeit im Januar also Winter und der höchste (Aphel) im Sommer datiert ist. Schau dazu mal [hier](#)<sup>1</sup> rein. Interessanterweise ändert sich dies aber über sehr lange Zeiträume. Es gibt nämlich verschiedene zyklische Einflüsse auf die Erde-Sonne-Kinematik die aus dem eigentlich einfachen Model ein sogar sehr kompliziertes und das Klima der Erde zu einem sehr dynamischen System machen. Dies spielt sich allerdings auf zum Teil sehr großen Zeitskalen ab. Falls es Dich interessiert, google mal nach MILANKOVIĆ-Zyklen bzw. gucke [hier](#)<sup>2</sup>.

---

1: [https://de.wikipedia.org/wiki/Apsis\\_\(Astronomie\)](https://de.wikipedia.org/wiki/Apsis_(Astronomie))  
2: <https://de.wikipedia.org/wiki/Milankovi%C4%87-Zyklen>



# Ergebnisinterpretation

Warum die zeitliche Verschiebung?

Aber es erklärt jetzt nur warum es im Sommer bei uns auf der Nordhalbkugel warm ist und im Winter kalt. Nicht aber den Versatz zwischen der mittleren täglichen Temperatur übers Jahr zur Sonnenscheindauer.

Die Sonnenscheindauer gehorcht genau der gleichen jährlichen Periode, wie auch die mittlere Tagestemperatur. Sie hängen alle beide gewissermaßen miteinander zusammen. Den größten Einfluss hat aber doch der Neigungswinkel. Nun ist dieser aber direkt korreliert mit der Sonnenscheindauer. Denn genau er bestimmt ja den Sinusverlauf der Sonne am Gestirn. Der „günstigste“ Neigungswinkel der Erdoberfläche ist genau an dem Tag, an dem auch die Sonne im Zenit steht und damit am längsten scheint.

Der Versatz der beiden Wetterparameter resultiert nun aus den unterschiedlichen Wärmekapazitäten von Luft und dem Boden. Die Luft kann wesentlich weniger Wärme speichern als der Boden. Die mittlere Tagestemperatur wird anhand der Lufttemperatur bestimmt. Und zwar über den gesamten Tag. Also nicht nur wenn die Sonne scheint, sondern eben auch nachts.

In den Wintermonaten, wenn das Bestrahlungspotenzial aufgrund des ungünstigen Neigungswinkel von Erde zur Sonne geringer ist als in den Sommermonaten und weniger lang die Sonne scheint, kühlst der Boden stark ab. Zwar kann sich die Luft tagsüber sehr schnell und teils sogar recht stark erwärmen (ich konnte mich auch so manches mal schon im Februar oben ohne sonnen), aber in der Nacht saugt der kalte Boden begierig die ohnehin geringe Menge an Wärme aus der Luft auf und erwärmt sich dabei allmählig. Man spricht von einem größeren Temperaturgradienten zwischen Boden und Luft. Je größer der nämlich ist, desto mehr Wärme kann pro Zeiteinheit vom Boden aus der Luft „gesaugt“ werden.

Betrachten wir nun den Zenit optimalsten Bestrahlung, also bei uns der 21. Juni, dann kann zwar die Sonnen am längsten auf den Boden scheinen, und auch am effektivsten bzgl. des Winkels, da der Boden ja der Sonne nun am optimalsten zugeneigt ist, aber er ist immer noch etwas ausgekühlt vom Winter und entzieht nachts daher immer noch etwas Wärme, was sich eben in einer verhältnismäßig geringeren Tagesdurchschnittstemperatur wiederspiegelt, verglichen mit der maximal wärmsten Zeit. Nach dem zeitlichen Zenit der Bestrahlungsdauer, also nach dem 21. Juni verschlechtert sich der Bestrahlungswinkel von Boden zur Sonne zunehmend wieder. Zudem nimmt eben gleichermaßen auch die Sonnenscheindauer wieder ab. Dennoch erreicht nun erst die mittlere Tagestemperatur ihren Zenit, weil nun der Boden in der Nacht, die gespeicherte Wärme aus Zeiten der optimaleren Bestrahlung an die Luft abgibt. Die Luft kühlst quasi etwas langsamer aus, als in den Tagen vor dem Erreichen des Zenits, obwohl Neigung der Erde und Tageslänge die gleichen sind, da der Temperaturgradient zwischen Luft und Boden nun geringer ist.

Es ist also ein Zusammenspiel zwischen Tageslänge, Bestrahlungswinkel, dem Breitengrad der Erde sowie dem Vermögen der Wärmespeicherung und Wärmeangabe des Bodens als auch der Luft.

Hierzu gibt es übrigens noch eine ganz [tolle Erklärung](#)<sup>1</sup> (auch in Ton) von GÁBOR PAÁL, veröffentlicht auf der Seite “Wissen” vom SWR.

1: <https://www.swr.de/wissen/1000-antworten/umwelt-und-natur/warum-ist-es-im-juli-und-august-am-heisstenen-obwohl-die-sonne-im-juni-am-hoechsten-steht-100.html>



# Schlusswort

*Ich hoffe ich konnte Dich mit dieser Arbeit verständlich durch mein Datenanalyseprojekt führen und zudem auch den Zugrunde liegenden Wettereffekt vernünftig erklären, ohne Dich dabei aber zu sehr mit komplizierter Mathematik und Physik konfrontiert zu haben. Ganz außer Acht gelassen habe ich sie ja nicht und sie zumindest im Groben mit auf den Weg gebracht. Wir konnten anhand von „rudimentären“ Wetterdaten einen astronomischen Prozess nachweisen. Natürlich sind dies keine neuen Erkenntnisse, aber hättest Du gedacht, dass man aus den Temperaturdaten diesen Effekt lesen kann?*

*Ich für meinen Teil hatte jedenfalls sehr viel Spaß bei dieser Arbeit und muss ehrlich sagen, dass ich Gefallen daran gefunden habe, populärwissenschaftlich zu schreiben. Ich bin sogar durch das Schreiben und die Analyse auf die Wetterdaten animiert wurden, noch eine weitere Arbeit über die Erderwärmung zu machen.*

*Leider sind die Ergebnisse in dieser Arbeit dann bei Weitem weniger schön wie die in diesem Projekt hier... nicht weil sie falsch sind, sondern leider genau das wiederspiegeln, was wir ja bereits längst wissen und leider zu oft verdrängen/missachten/ignorieren.... Mit anderen Worten, ich konnte aus genau diesen Wetterdaten der mittleren Tagstemperatur auch die globale Erderwärmung analysieren. Es wird darauf also auch eine kleine Arbeit wie diese hier folgen... Nur dann leider mit etwas weniger Witz und dafür umso mehr Ernst.*

*Falls Dir die Arbeit gefallen hat, freue ich mich über Deine Kritik!*

*Natürlich sind auch Anregungen gern gesehen. Und zögere nicht, mich freundlich auf Fehler hinzuweisen. Ich bin mir sicher welche, trotz größtem Versuch der Mühe und Sorgfalt, gemacht zu haben.*

*Verteile dieses Werk bitte gerne an möglicherweise Interessierte weiter! Ich möchte einfach ein paar Leute damit erreichen. Nicht weil ich mich mit Rum und Kohle bekleckern will - die Arbeit ist ja eh für Lau, sondern, weil ich versuchen möchte so viele Leute wie möglich für Datenanalyse und Naturwissenschaften zu begeistern.*

*Viele Grüße,*

*Alex*

# Danksagung und Erwähnungen

An erster Stelle danke ich meinen Stiefvater DIPL.-ING. GÖTZ WEBER für die klasse Tricks bzgl. der internen Verlinkungen, um dieser Arbeit die gewisse Interaktivität zu verleihen.

~ ~ ~

Mathematiker haben oft einen schlechten Ruf. Insbesondere unter Schülern und Studenten. Nämlich den, dass sie Sachverhalte nicht genügend „gut“ erklären können.

Mag sein, dass das sehr subjektiv ist, aber im Großen und Ganzen wird das schon auf so einige – vermutlich sogar die Meisten – zutreffen.

Aber es gibt sie, die Ausnahmen in dieser kleinen subjektiven Statistik. Ich möchte hier 4 sehr bekannte Namen erwähnen, welche mich durch meine Studienjahre insbesondere eben den zahlreichen Mathematik- und auch Physikklausuren durch didaktisch optimal designete Wissensvermittlung in Form ihrer Youtube-Videos gebracht haben:

PROF. DR. CHRISTIAN SPANNAGEL

PROF. DR. JÖRN LOVISCACH

PROF. DR. EDMUND WEITZ      (...ich liebe die Weihnachtsvorlesungen!)

DANIEL JUNG                    (...dessen akademischen Grad ich leider nicht kenne)

Herausragende Mathematiker vor allem deshalb, weil sie Mathematik anschaulich und verständlich vermitteln. Vielen Dank ihnen dafür!!!

Ich danke:

PROF. DR. STEFAN SCHLECHTWEG als Dozent des Moduls „Informationssvisualisierung und Visual Analytics“ an der Hochschule Anhalt für seinen klasse Lehrstil, der sehr angenehmen Art, dem vielen vermittelten Wissen, sowie dem „Muse“ dieses Projekts sein.

Dem Deutschen Wetter Dienst DWD für die freie Bereitstellung ihrer Wetterdaten.