

## Praktikum Smart Data Analytics, Team 4 – 2. Übungsblatt

### Inhalt

<b>Praktikum Smart Data Analytics, Team 4 – 2. Übungsblatt</b>	1
<b>Einleitung</b>	2
<b>Temperaturtrends in Deutschland</b>	2
Vorbereitung	2
Bundes- und Bundesländer-Ebene	3
Landkreisebene	10
Erklärungsmöglichkeiten für Unterschiede	12
<b>Lokale Unterschiede in den Trends: Baden-Württemberg</b>	14
<b>Clustern von Temperaturverläufen: Baden-Württemberg</b>	15
Clustering mit Korrelation	16
Clustering mit Cosine-Distance	18
<b>Analyse der relevanten Cluster</b>	19
Zeitliche Verläufe der Cluster	19
Geographischen Gesichtspunkte der Cluster	24
<b>Alternative Clusteringverfahren</b>	25
<b>Bedeutung der Datenqualität und Metainformationen</b>	26
<b>Zusammenfassung und Ausblick</b>	28

## Einleitung

Das Klima und seine Entwicklung beeinflussen das menschliche Leben auf der Erde maßgeblich. Daher ist insbesondere der globale Klimawandel ein Phänomen, das große wissenschaftliche und politische Aufmerksamkeit auf sich zieht und eines der wichtigsten Themen unserer Zeit darstellt.

Doch wie entwickeln sich Klima und Wetter regional oder sogar lokal? Welche Trends gibt es in Deutschland und wie unterschiedlich sind die Regionen betroffen? In diesem Bericht möchten wir mithilfe von Data-Science-Verfahren und ihren Werkzeugen explorieren, welche Erkenntnisse aus der temporal-geografischen Untersuchung von Lufttemperaturdaten in Deutschland gewonnen werden können. Darüber hinaus betrachten und diskutieren wir, was diese Erkenntnisse bedeuten und welche Konsequenzen daraus folgen können.

## Temperaturtrends in Deutschland

Wie gleichmäßig und unterschiedlich entwickeln sich die Temperaturen in Deutschland?

### Vorbereitung

Bevor wir mit der Untersuchung des DWD-Datensatzes Lufttemperatur 2010-2019 hinsichtlich der Temperaturentwicklung in Deutschland beginnen können, müssen wir uns mit dessen Struktur auseinandersetzen und näher betrachten, wie er erhoben wurde. Hierzu visualisieren wir in einem ersten Explorationsschritt die Zahl der insgesamt 532 Messstationen im Jahr 2019 im Kontext der politischen Regionen Deutschlands. Dies realisieren wir neben Recherchen mit *Cadenza* in einem Python-Skript, das Geopandas-Datenstrukturen und die *Folium*-Kartenbibliothek mit *OpenStreetmap*-Daten verwendet. Wir setzen geografische Beschreibungen der Bundesländer und Landkreise Deutschlands im Geo-Json-Format.

Um eine Zuordnung zu kleineren politischen Einheiten zu erhalten, führen wir zuerst eine Zuordnung von Messstationen zu Landkreisen durch. Dies gelingt durch den Einsatz von Geopandas-Datenstrukturen mit integrierten geometrischen Schnitttests und der geografischen Beschreibung der Landkreise in Deutschland. Jeder Messstation wird genau der eindeutige Name des Landkreises zugeordnet, in dem sie sich befindet.

Abbildung 1 zeigt die Visualisierung der Messstationen.

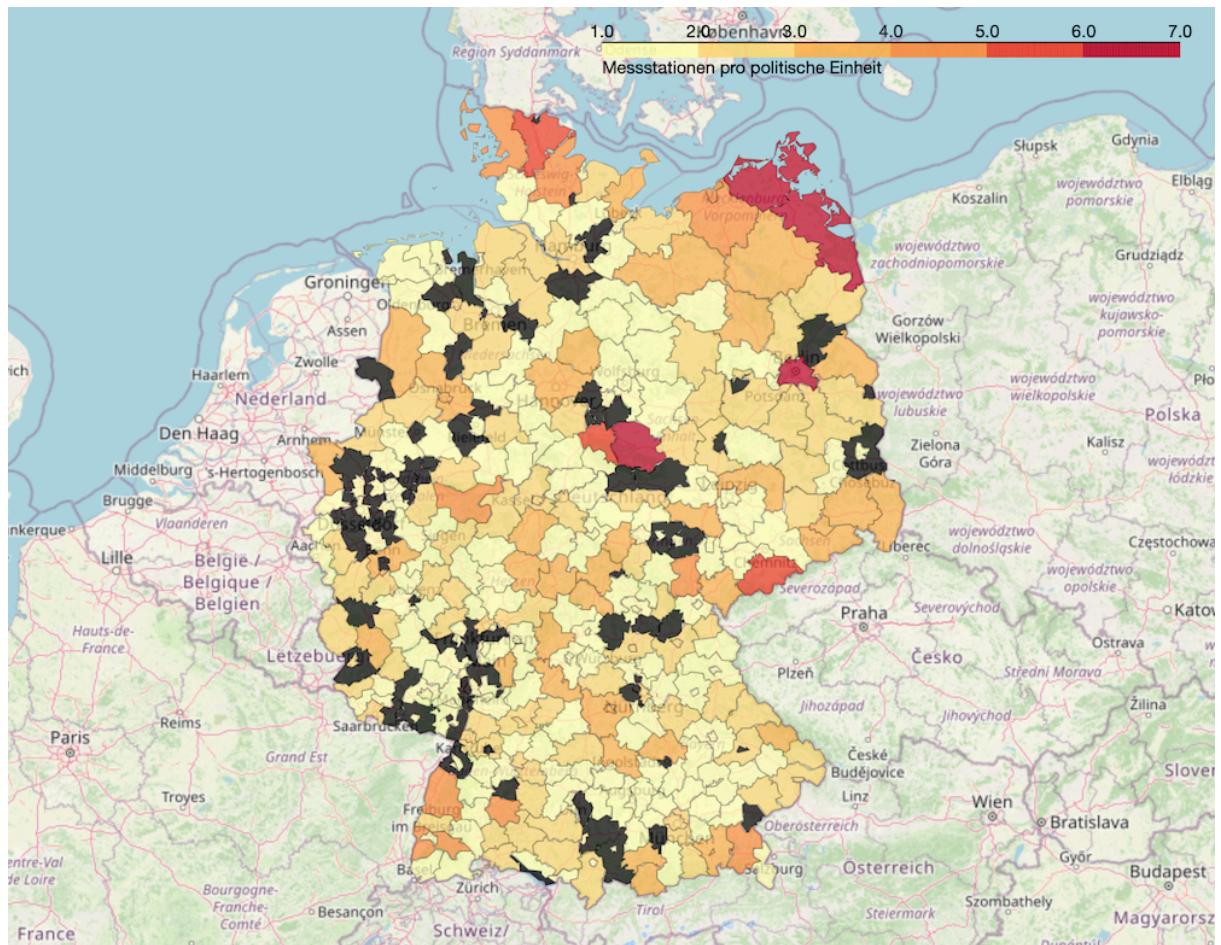


Abbildung 1: Visualisierung der Anzahl der Lufttemperatur-Messstationsanzahl in den Landkreisen Deutschlands.

Auf Ebene der Landkreise ist sofort erkennbar, dass fast alle Regionen über mindestens eine, jedoch nur wenige Kreise über mehr als vier Messstationen verfügen. Beispielsweise im Nordosten entlang der Küste ist die Dichte der Stationen hoch. Einige Regionen beinhalten keinerlei Messstation für die Lufttemperatur (in der Abbildung schwarz dargestellt), jedoch gibt es in fast allen Fällen eine nahegelegene Station in angrenzenden Kreisen, so dass dies keine Aussage über die Qualität der Messungen bezüglich der politischen Kreise ist, sondern lediglich eine grundlegende organisatorische Betrachtung. Kein Landkreis beinhaltet mehr als sieben Stationen.

#### Bundes- und Bundesländer-Ebene

Unser Analysegegenstand ist die temporale Entwicklung der Lufttemperaturen. Zuerst wollen wir den Temperaturverlauf in den Bundesländern Deutschlands darstellen. Hierzu erstellen wir einen Plot aller

Verläufe der Durchschnitts-Jahrestemperaturen, getrennt nach den Bundesländern. Es kommen wieder Python und Geopandas zum Einsatz, nun mit einer Matplot-Visualisierung der Zeitserien.

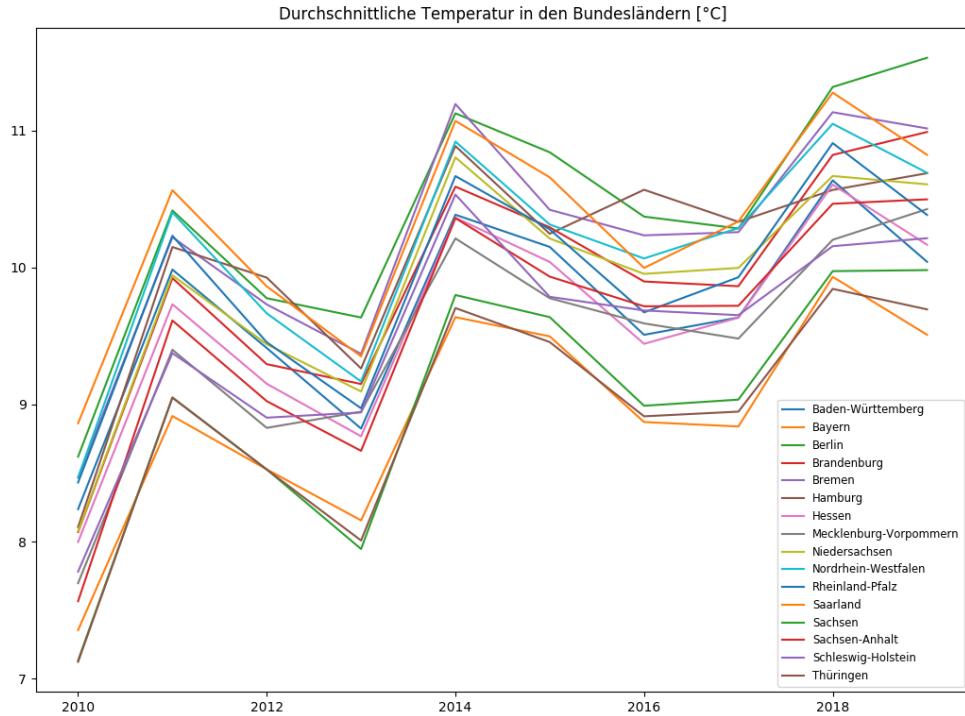


Abbildung 2: Visualisierung der Verläufe der durchschnittlichen Jahres-Lufttemperatur der Bundesländer Deutschlands, 2010-2019.

Qualitativ kann aus Abbildung 2 abgelesen werden, dass die durchschnittliche Jahrestemperatur vieler Bundesländer zueinander positiv korreliert ist und die Durchschnittstemperatur in der Tendenz im Betrachtungszeitraum zunimmt.

Um temporale Effekte in der Temperaturentwicklung qualitativ auch geografisch zu überblicken, fertigen wir eine Animation der Temperaturentwicklungen mithilfe der Python-Bibliothek Folium und dem Videoverarbeitungs-Toolkit *FFMPEG* an, die den gesamten Zeitraum von 2010 bis 2019 auf Ebene der Bundesländer visualisiert. Hierbei werden zuerst wieder Durchschnittstemperaturen über die Jahre verwendet.

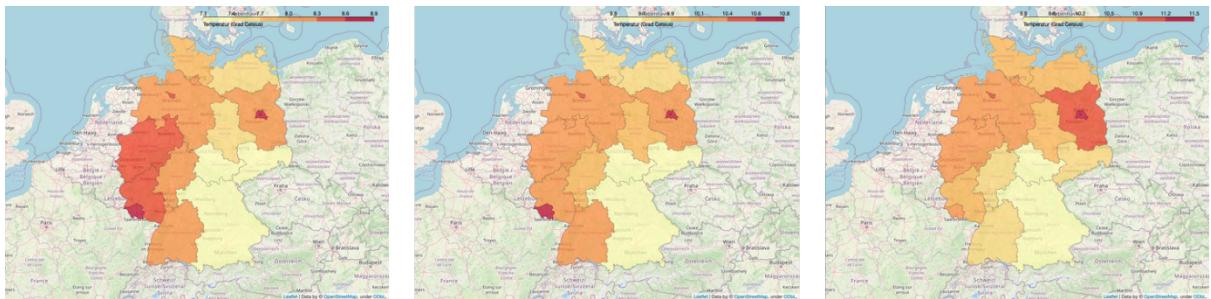


Abbildung 3: Standbilder der temporalen Visualisierung der durchschnittlichen Lufttemperaturen in den Bundesländern. Hier dargestellt sind die Jahre 2010 (links), 2015 (Mitte) und 2019 (rechts).

Es werden schnell drei Regionen sichtbar, die als rot eingefärbt regelmäßig die wärmsten durchschnittlichen Lufttemperaturen aufweisen: Berlin, Hamburg und das Saarland. Diese qualitative Erkenntnis wollen wir später quantitativ genauer betrachten. Grundsätzlich zeigt der Nordosten, Norden und Westen Deutschlands in der Visualisierung die höchsten Durchschnittstemperaturen.

Wir bereiten die Daten für die Bundesländer zunächst mit statistischen Mitteln weiter auf und stellen sie kompakter dar, um einen besseren Überblick zu erhalten.

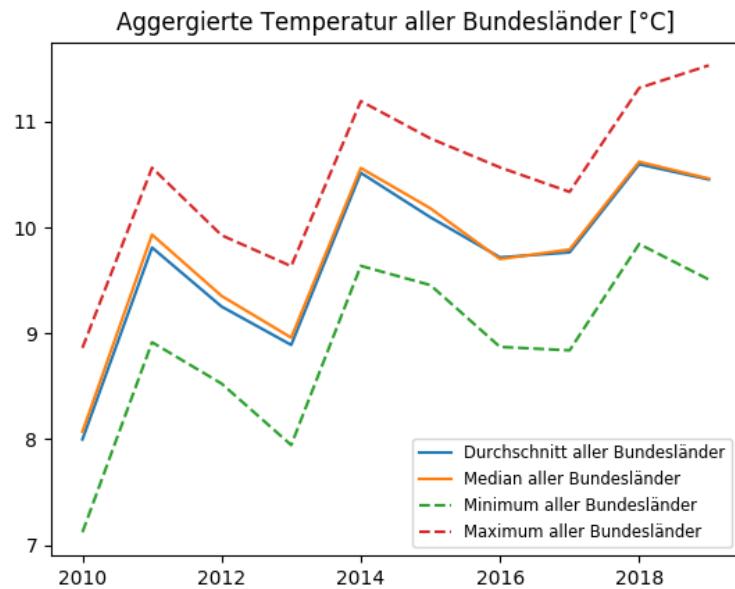
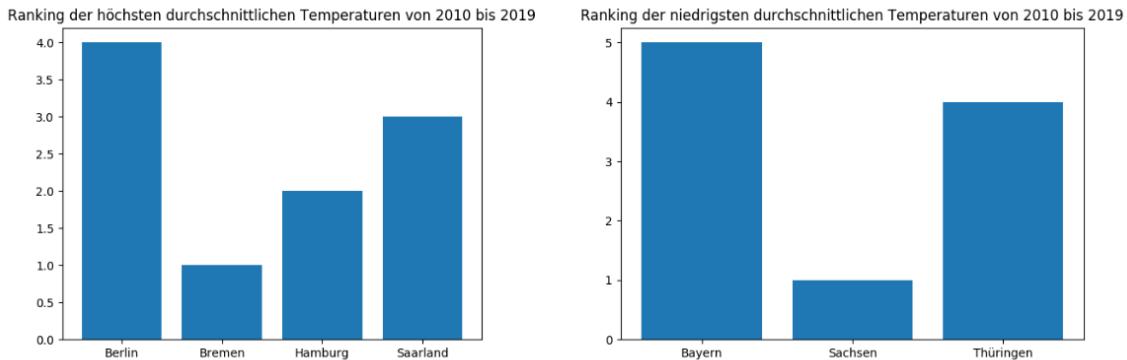


Abbildung 4: Aggregierte Durchschnittstemperaturen der deutschen Bundesländer als Zeitserien auf Jahresentebene. Mittelwert und Median liegen im Betrachtungszeitraum nahe beieinander, Minimum und Maximum der Länder-Durchschnittstemperaturen liegen etwa symmetrisch um den Mittelwert.

Mithilfe des Durchschnitts der Zeitserien des ersten Plots bestätigt sich der Eindruck, dass ein grundsätzlicher Anstieg der Durchschnittstemperatur im Bundesgebiet im Betrachtungszeitraum vorliegt. Der Median entspricht fast genau dem Verlauf des Durchschnitts über allen Bundesländern, somit liegt in etwa die Hälfte aller Bundesländer über und die andere Hälfte unter der durchschnittlichen Temperaturentwicklung. Minimum und Maximum der Jahres-Durchschnittswerte der Lufttemperaturen in den Bundesländern liegen in etwa symmetrisch um Mittelwert und Median und liegen in einer Streuung von etwa einem Grad Celsius um diesen Durchschnitt. Zwischen Minimum und Maximum der jährlichen Durchschnittswerte der Bundesländer liegen im Betrachtungszeitraum also stets in etwa 2°C. Zwischen 2010 und 2019 hat die durchschnittliche Jahrestemperatur in Deutschland um etwa 2,4°C zugenommen.

Wir erstellen zur weiteren Untersuchung ein Ranking der durchschnittlichen jährlichen Lufttemperaturen in den Bundesländern. Hierbei interessiert uns, wie oft Länder die im Jahresdurchschnitt höchste und wie oft die niedrigste Temperatur im Bundesgebiet vorweisen. Wir stellen die Anzahl der Platzierung auf dem ersten bzw. letzten Position des Rankings im Folgenden in Diagrammen dar (Abbildung 5).

Im Betrachtungszeitraum wies Berlin viermal, das Saarland dreimal, Hamburg zweimal und einmal



*Abbildung 5: Darstellung der Anzahl der Belegungen des ersten Platzes der höchsten (links) und des ersten Platzes der niedrigsten Durchschnittstemperaturen (rechts) im Ranking der Bundesländer, als Balkendiagramme. Im Betrachtungszeitraum waren Berlin, das Saarland, Hamburg und Bremen am öftesten die im Jahresschnitt wärmsten Bundesländer. Bayern, Sachsen und Thüringen waren am öftesten die im Jahresmittelwert kältesten Länder.*

Bremen die höchste jährliche Durchschnittstemperatur in Deutschland auf. Insbesondere handelt es sich hierbei um eher kleine Bundesländer, drei von ihnen sind Stadtstaaten. Die niedrigsten Durchschnittstemperaturen finden wir in Bayern (fünfmal), viermal in Thüringen und einmal in Sachsen. Insgesamt sind Berlin und das Saarland also die wärmsten, Bayern und Thüringen im Betrachtungszeitraum die im Durchschnitt kältesten Bundesländer.

#### *Saisonalität*

Der Jahresmittelwert der Lufttemperaturen ist ein sehr stark aggregierendes Werkzeug. Wir möchten nun die Temperaturentwicklung in den Jahreszeiten Sommer (eingeschränkt auf Juni, Juli, August) und Winter (November, Dezember, Januar) herausgreifen und getrennt betrachten. Zuerst plotten wir auch für diesen wieder Übersichten, um die Entwicklung der Zeitserien der Durchschnittstemperaturen besser zu überblicken.

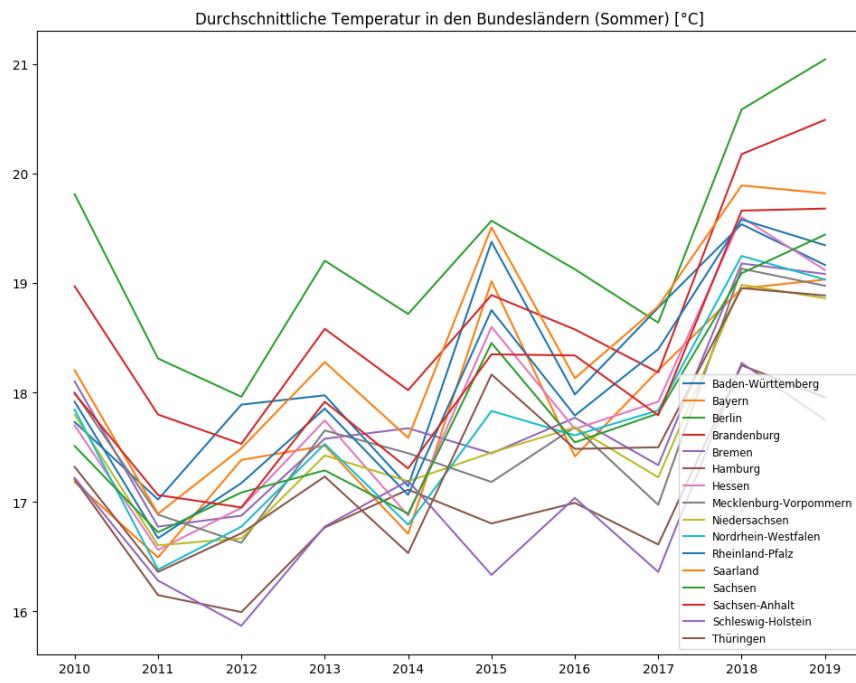


Abbildung 6: Visualisierung der Verläufe der durchschnittlichen Lufttemperatur im Sommer der Bundesländer Deutschlands nach Jahren, 2010-2019.

Es lassen sich bereits in Abbildung 6 eine Reihe interessanter Aspekte ablesen, wie Beispielsweise, dass Berlin auch beschränkt auf die Sommermonate meiste die wärmsten Durchschnittstemperaturen aufweist und dass ein grundsätzlich ansteigender Trend über die Bundesländer im Betrachtungszeitraum zu erkennen ist. Wir setzen wieder statistische Mittel ein, um die Trends übersichtlicher darzustellen (vgl. Abbildung 7).

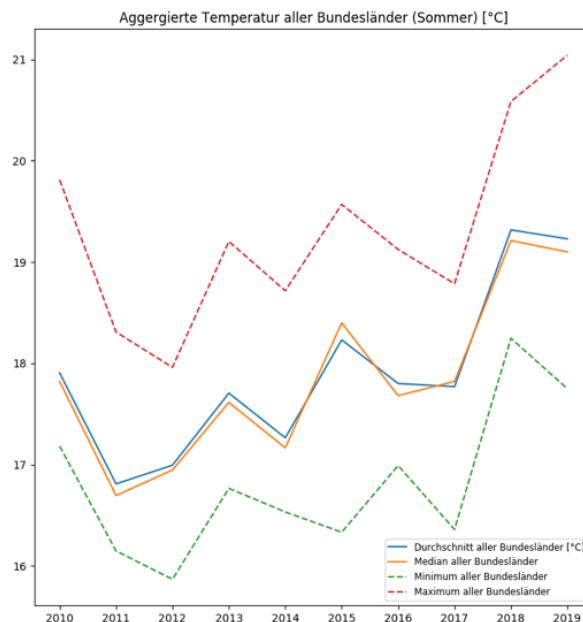


Abbildung 7: Aggregierte Durchschnittstemperaturen der deutschen Bundesländer im Sommer als Zeitserien auf Jahresschicht.

Hier wird deutlich, dass die jährliche Streuung der Mittelwerte im Sommer teils weiter um den Durchschnitt und Median herum liegt, als in der ganzjährlichen Betrachtung. Mittelwert- und Median-Zeitserien über die Sommer-Mittelwerte der Bundesländer sind auch hier fast identisch. Grundsätzlich unterscheiden sich die Zeitserien jedoch stark zur ganzjährlichen Betrachtung.

Wir erstellen dieselben Plots ebenfalls für die zuvor festgelegten Wintermonate.

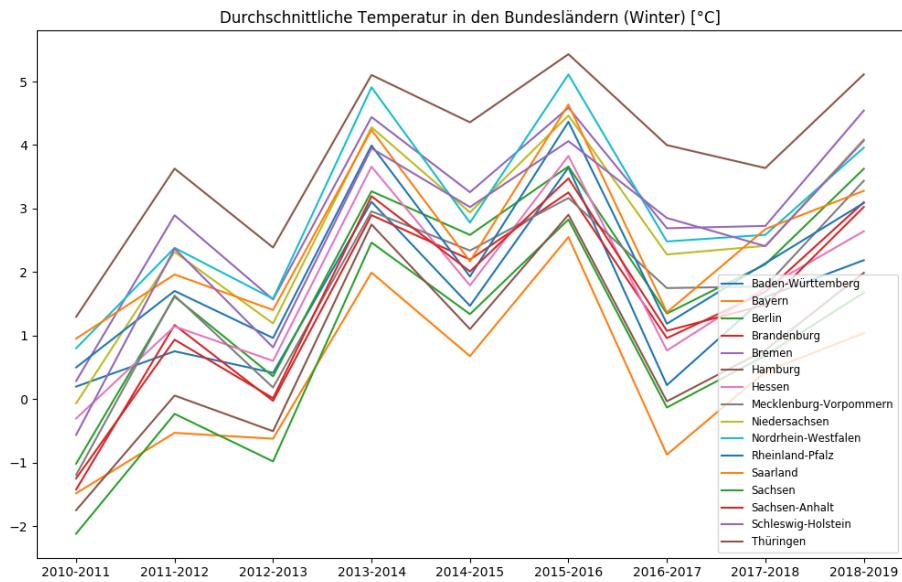
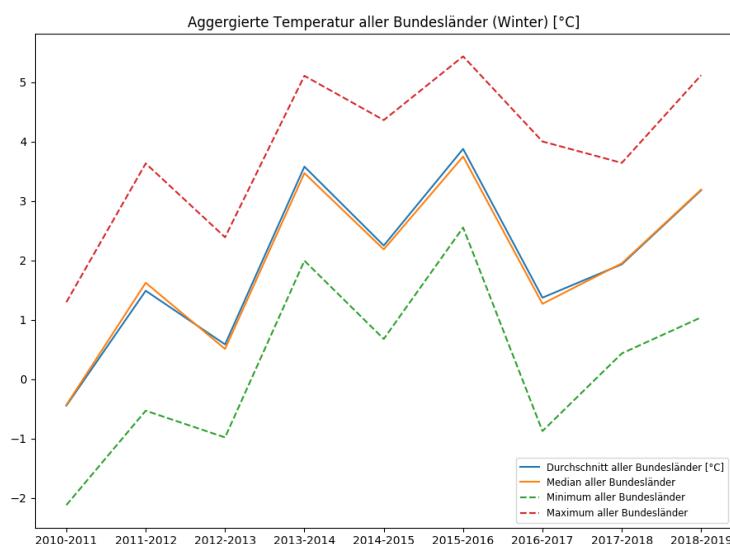


Abbildung 8: Visualisierung der Verläufe der durchschnittlichen Lufttemperatur im Winter der Bundesländer Deutschlands nach Jahren, 2010-2019.

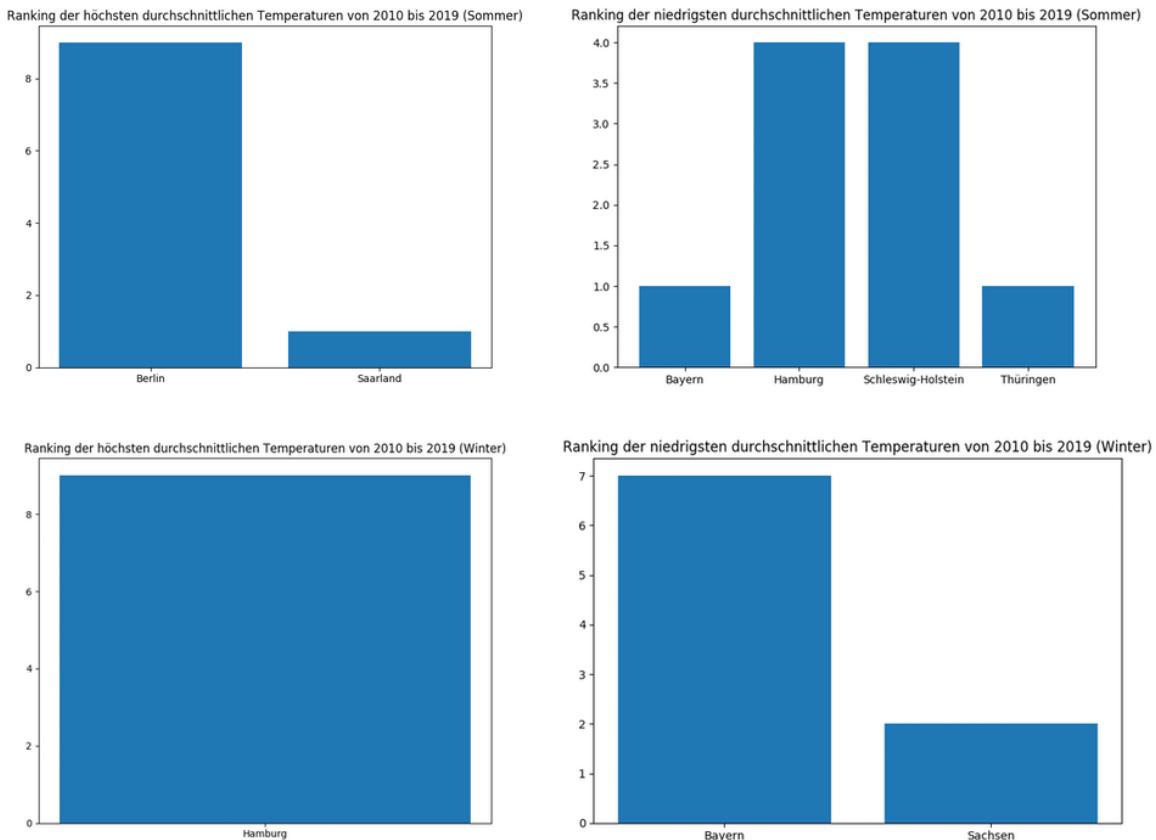
In den Wintermonaten können wir aus dem Plot sofort Hamburg als wärmstes Bundesland ablesen, dies ist für alle Jahre im Betrachtungszeitraum der Fall.



*Abbildung 9: Aggregierte Durchschnittstemperaturen der deutschen Bundesländer im Winter als Zeitserien auf Jahresebene.*

Die statistischen Größen im Winter zeigen uns ein erwartetes Bild, da wir sie für das ganze Jahr und die Sommermonate bereits kennen. Auch hier sind Mittelwerte und Mediane der Durchschnittstemperaturen fast identisch. Allerdings ist die Streuung der Temperaturen um den Mittelwert deutlich höher, teils bis zu 2°C nach oben und unten.

Nun wollen wir wieder, wie für das ganze Jahr bereits erstellt, ein Ranking der am häufigsten wärmsten und der am häufigsten kältesten Bundesländer betrachten. Abbildung 10 stellt die Ergebnisse der Zählung der Erstplatzierungen als Balkendiagramme dar.



*Abbildung 10: Darstellung der Anzahl der Belegungen des ersten Platzes der höchsten (links) und des ersten Platzes der niedrigsten Durchschnittstemperaturen (rechts) im Ranking der Bundesländer, als Balkendiagramme. In dieser Analyse getrennt betrachtet eingeschränkt auf Sommer (oben) und Winter (unten).*

Im gesamtem Betrachtungszeitraum liegen die höchsten Sommertemperaturen in Berlin (neunmal) und einmal im Saarland vor. Die kältesten durchschnittlichen Temperaturen im Sommer zeigen Hamburg und Schleswig-Holstein (beide viermal) und je einmal Bayern und Thüringen. Im Winter hingegen, wie schon in Abbildung 8 abzulesen war, sind die höchsten Durchschnittstemperaturen stets in Hamburg zu finden. Bayern ist siebenmal das im Durchschnitt kälteste Bundesland im Winter, gefolgt von Sachsen (zweimal) im Betrachtungszeitraum.

### Änderungsrate

Nun möchten wir die Änderungsrate der Temperaturentwicklung in den Bundesländern betrachten. Hierzu ziehen wir wieder die ganzjährlichen Durchschnittswerte der Bundesländern heran und berechnen die jährliche Änderung. Abbildung 11 zeigt eine Visualisierung der Änderungs-Zeitserien.

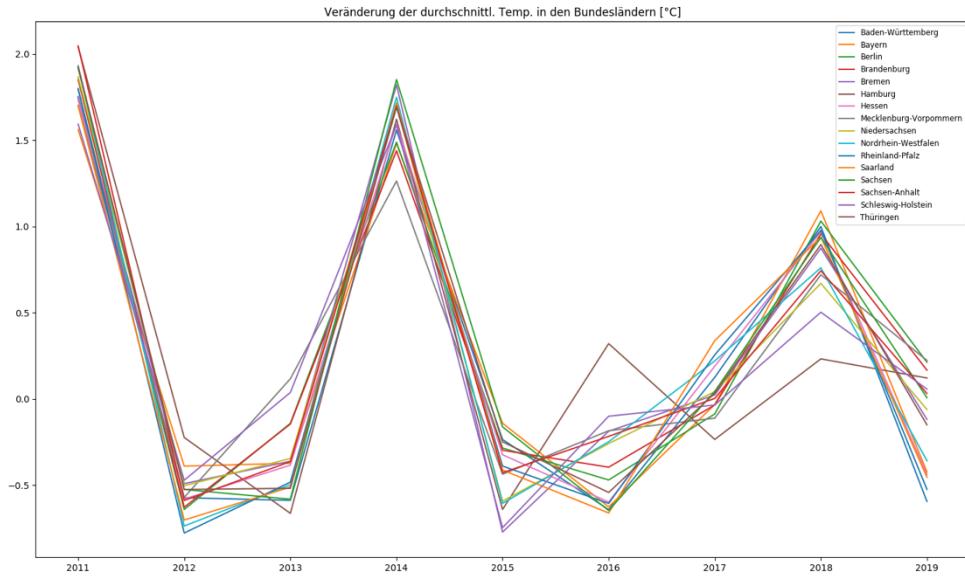


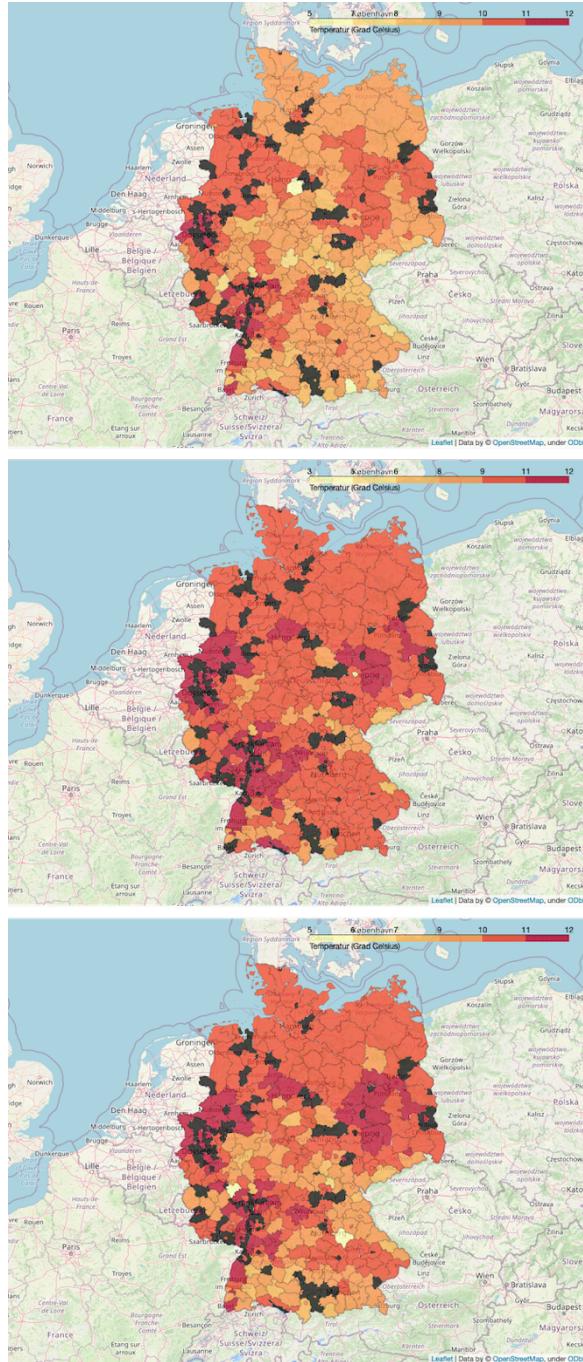
Abbildung 11: Zeitserien der jährlichen Änderungsrate der Durchschnittstemperatur in den Bundesländern in Deutschland.

Aus dem Plot können wir ablesen, dass die jährliche Durchschnittstemperatur in den Jahren 2011, 2014 und 2018 bundesweit am stärksten zunahm und in den übrigen Jahren nur leicht zurückging. Die Varianz der Änderung zwischen den Bundesländern ist bis auf wenige Ausreißer gering. Beispielsweise nahm vom Jahr 2015 auf das Jahr 2016 die durchschnittliche Lufttemperatur in Hamburg um fast 0,4°C zu, während alle anderen Bundesländer einen leichten Rückgang der Durchschnittstemperaturen verzeichneten.

### Landkreisebene

Da wir in der Vorbereitung unserer Analyse bereits eine Zuordnung von Messstationen zu den Landkreisen angefertigt haben, können wir diese Informationen nutzen, um auch Temperaturtrends auf regionaler Ebene zu visualisieren. Hierbei nehmen wir keine Interpolation für Landkreise ohne Messstationen vor, sondern interessieren und lediglich für allgemeine Trends, die auch mit der großen Anzahl der übrigens Kreise sichtbar werden.

Wir erstellen mithilfe von Folium und FFMPEG wieder temporal-geografische Darstellungen in Form von Animationen, um Qualitative Aussagen zu erhalten.



*Abbildung 12: Standbilder der temporalen Visualisierung der durchschnittlichen Lufttemperaturen in den Landkreisen. Hier für die Jahre 2011 (oben), 2016 (Mitte) und 2019 (unten).*

Auch hier lässt sich erkennen, dass die im Durchschnitt wärmsten Temperaturen im Norden und Westen Deutschlands auftreten. Außerdem lassen sich warme Gebiete wie die Oberrheinische Tiefebene im Südwesten Deutschlands erkennen, die durch Mittelwertbildung innerhalb der Bundesländer vorher nicht sichtbar waren. Weiterhin stechen kleine Regionen wie das Saarland, Berlin oder Hamburg hervor, hier wird jedoch auch sichtbar, dass andere Bundesländer ebenfalls über

kleinere im Durchschnitt wärmere Regionen verfügen, die bisher nur nicht separat betrachtet wurden und somit in der Aggregation über die größeren Bundesländer verschwanden.

### *Ballungszentren*

Wir möchten in unserer Analyse nun eine Trennung der Landkreise in die Ballungszentren Deutschlands und die übrigen, weniger dicht besiedelte Kreise vornehmen.

Hierzu klassifizieren wir die Landkreise in urbane Regionen und sonstige Kreise, entlang der Klassifikationsschwelle von 500.000 Einwohnern im Jahr 2018. Als signifikante Großstädte und Ballungszentren betrachten wir demnach die Kreise:

Berlin, Hamburg, München, Köln, Frankfurt am Main, Stuttgart, Düsseldorf, Dortmund, Essen, Bremen, Region Hannover, Leipzig, Dresden, Nürnberg, und Duisburg.

In einer Untersuchung der durchschnittlichen Jahrestemperaturen im Vergleich zwischen diesen Ballungszentren und anderen Landkreisen lässt sich jedoch kein Unterschied feststellen. Das Kriterium Ballungszentrum ist im DWD-Messnetz für die Lufttemperatur also kein Merkmal, das alleine hinreichend ist um einen auffälligen Trend bei der Temperaturentwicklung zu identifizieren.

Eine analoge Analyse, eingeschränkt auf die größten Städte in Baden-Württemberg (Stuttgart, Karlsruhe, Mannheim, Freiburg im Breisgau und Heidelberg) zeigt ebenfalls keine signifikante Abweichung der Städte im Vergleich zu den anderen Landkreisen im Bundesland. Somit ist auch im Bundesland nicht von einem Trend auszugehen, der sich alleine durch die Bevölkerungsdichte von anderen Regionen unterscheidet.

### *Erklärungsmöglichkeiten für Unterschiede*

In Deutschland herrscht ein Übergangsbereich zwischen maritimen (küstennahem) und kontinentalem Wetter (Quelle). Das kann dazu führen, dass in unterschiedlichen Bereichen in Deutschland deutlich unterschiedliche Temperaturen vorzufinden sind.

### *Unterschiede West/Ost*

Beispielsweise kann es zu Abweichungen zwischen West und Ost-Deutschland kommen. Durch die Westwinde strömt häufig milde Meeresluft vom Atlantik nach Deutschland. Da der Einfluss der milden Meeresluft von West nach Ost-Deutschland abnimmt, dominiert im Westen vorwiegend das maritime Klima, während im Osten Deutschlands ein verstärkt kontinentales Klima vorzufinden ist. Dies spiegelt sich in den vorangegangen Analysen wider, da in westlichen Wetterstationen geringere Abweichungen zwischen Sommer- und Wintermonaten vorzufinden sind als in den östlichen (Datenbezug, am besten mit Grafik). Ein maritimes Klima zeichnet sich durch geringere Temperaturunterschiede zwischen Sommer und Winter aus.

### *Unterschiede Nord/Süd*

Im Nord-Süd Vergleich lässt sich ebenfalls der Übergang zwischen maritimen und kontinentalem Wetter als Erklärungsmöglichkeit für die identifizierten Abweichungen hinzuziehen. Im Norden Deutschlands sorgt die Wassernähe zu einer höheren Feuchte in der Atmosphäre. Im Süden Deutschlands hingegen, ist das Klima wesentlich kontinentaler, was bedeutet, dass die Luft trockener ist. Ebenfalls zeigen weitere Analysen der DWD Daten, dass im Süden Deutschlands mehr Städte mit

vielen Sommertagen liegen<sup>1</sup>. Ein Sommertag in diesem Kontext beschreibt einen Tag mit mehr als 25 Grad Celsius und einem Himmel der mindestens zur Hälfte frei von Wolken ist.

#### *Bundeslandebene*

Die beschriebenen Unterschiede auf Bundeslandebene in der vorangegangenen Datenanalyse lassen sich ebenfalls aufgrund geographischer Gegebenheiten erklären. Beispielsweise kommt es zu höheren Durchschnittstemperaturen bei den Wetterstationen in Baden-Württemberg, da viele der Stationen in der oberrheinischen Tiefebene liegen, die durch ihre natürliche Senke eine einfache Aufheizung aber erschwert Abkühlung der Temperaturen ermöglicht. Betrachtet man nur die gemessenen Werte der Wetterstationen, die in der oberrheinischen Tiefebene liegen (Abbildung 13), erhalten wir eine Durchschnittstemperatur von 10.9°C über den gesamten Messungsverlauf, was deutlich über dem Durchschnitt von BW liegt (9.7°C) und auch über dem Gesamtdurchschnitt (9.5°C).

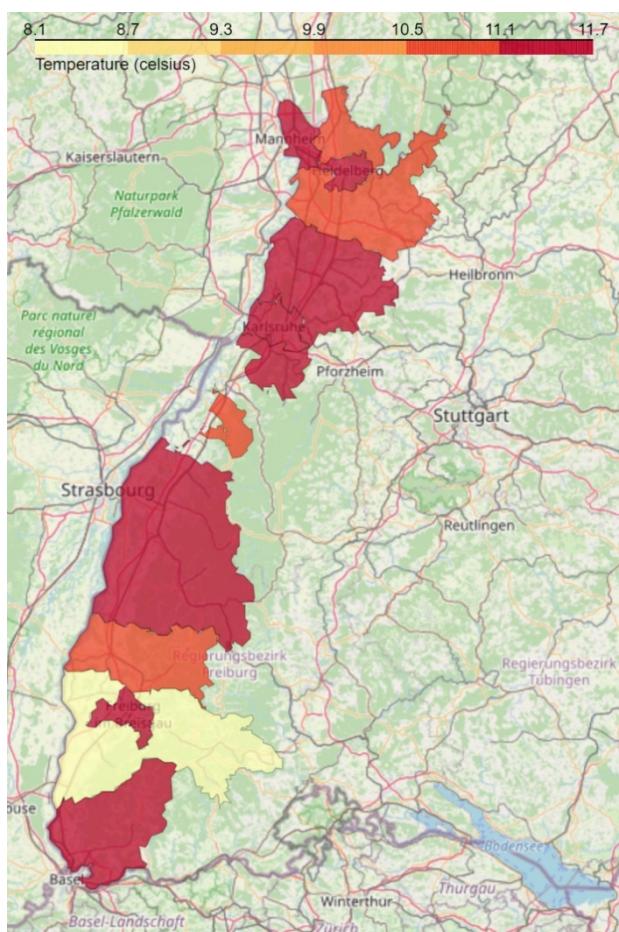


Abbildung 13: Durchschnittstemperaturen aggregiert auf Landkreisebene für Wetterstationen in der Oberrheinischen Tiefebene

Das Ranking der niedrigsten bzw. der höchsten durchschnittlichen Temperaturen (Abbildung 10) bestätigt uns am Beispiel Hamburg das vorherrschende maritime Klima. Aufgrund der Wassernähe verzeichnet Hamburg so oft wie kein anderes Bundesland die kältesten Sommer bzw. wärmsten Winter.

---

<sup>1</sup> <https://www.sueddeutsche.de/wissen/wetterdaten-analyse-hier-kommt-der-sommer-1.3143843>

Auffällig ist auch, dass ausschließlich kleinere Bundesländer (Berlin, Bremen, Hamburg, Saarland) die höchsten Jahresdurchschnittstemperaturen erzielen.

### Lokale Unterschiede in den Trends: Baden-Württemberg

Um mögliche Ausreißer zu identifizieren betrachten wir zunächst ausschließlich die täglichen Temperaturwerte der einzelnen Wetterstationen in Baden-Württemberg (BW) und vergleichen diese mit der Durchschnittstemperatur des entsprechenden Tages von ganz BW. Wir berechnen die tägliche Differenz der einzelnen Stationen vom Tagesdurchschnittswert und betrachten konkret den Root-Mean-Squared-Error (RMSE) für jede Station. Dadurch ergeben sich 5 Wetterstationen, die sich durchschnittlich über den gesamten Zeitverlauf hinweg um mehr als 2.25 Grad von der täglichen Durchschnittstemperatur unterscheiden. Diese 5 Stationen sind in Abbildung 14 farblich hervorgehoben.

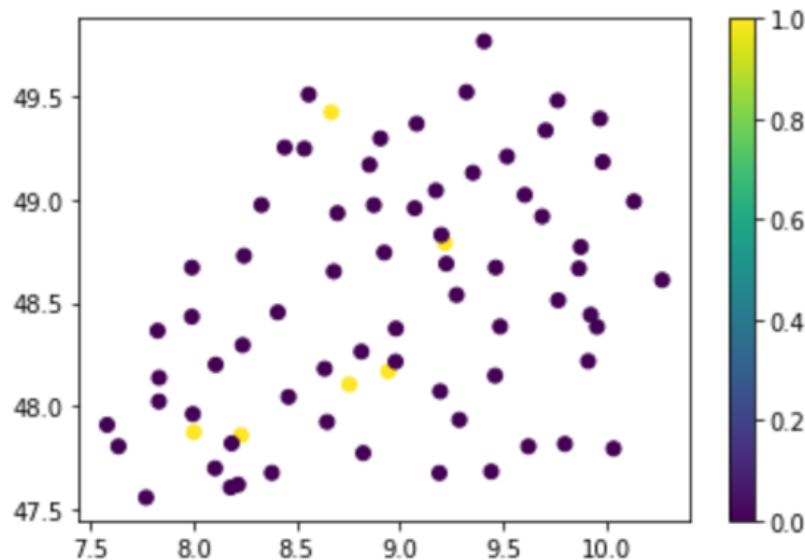


Abbildung 14: Wetterstationen mit größter durchschnittlicher Abweichung von Tagesdurchschnittswert.

Betrachtet man diese 5 Stationen genauer, erkennt man, dass diese Messstationen deutlich höher bzw. tiefer liegen als der Rest der Stationen und sich daher der Temperaturunterschied erklären lässt.

Um genauer zu untersuchen, ob und inwieweit einzelne Wetterstationen sich gegen den Trend entwickeln, betrachten wir die Korrelation zwischen deren Tageswerten der einzelnen Stationen und den Landesdurchschnitt. Erneut erkennen wir, dass vor allem Stationen in höheren Lagen dem Gesamtrend schwächer folgen (Korrelation geringer als 0.98).

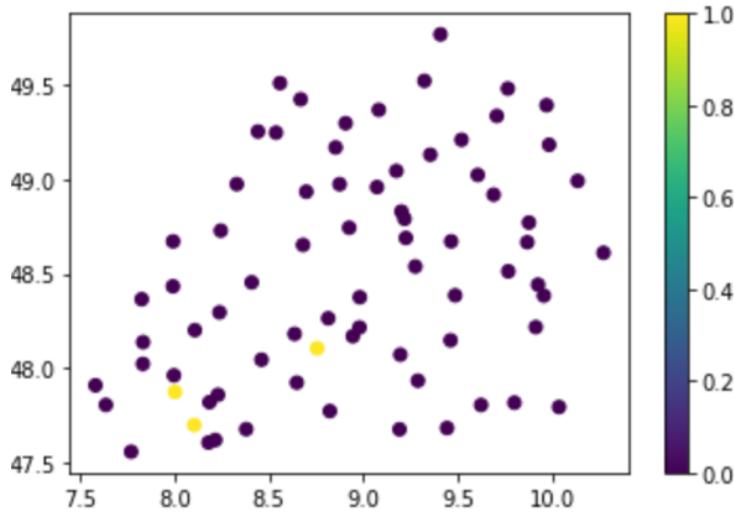


Abbildung 15: Wetterstationen, die unterdurchschnittlich mit Gesamtrend von BW korrelieren.

### Clustern von Temperaturverläufen: Baden-Württemberg

Wir werden nun versuchen, automatisch lokale Cluster von Stationen mit ähnlichen Temperaturverläufen zu identifizieren. Hierbei beschränken wir unsere Betrachtungen erneut auf die Region des Landes Baden-Württemberg.

Eine besondere Herausforderung stellt hierbei die Mischung von Orts- und Zeitbezogenen Daten dar. Wir benötigen also einen Clustering-Ansatz, der gleichzeitig sowohl die Ähnlichkeit der Temperaturverläufe als auch die geografische Nähe der einzelnen Stationen berücksichtigt.

Für jede Cluster-Analyse müssen zunächst zwei essenzielle Teilaufgaben festgelegt werden: Eine Distanzfunktion, die festlegt, welche Datenpunkte als „nahe zusammenliegend“ und welche als „weit entfernt“ betrachtet werden, sowie ein Clustering-Algorithmus, der dann auf Basis dieser Abstände ein Clustering erzeugt.

Aus der Vielzahl verschiedener Clusteringverfahren und -ansätze haben wir uns für das dichtebasierende DBSCAN-Verfahren entschieden. Essenziell für die Auswahl war die Bedingung, dass nicht-euklidische Distanzmetriken unterstützt werden müssen. Weiterhin besitzt DBSCAN folgende Vorteile:

- Es werden Cluster beliebiger Form erkannt (im Gegensatz z.B. zur Annahme von Oval förmiger Cluster bei anderen Ansätzen).
- Die Anzahl der Cluster muss nicht bereits im Vorhinein festgelegt werden (wie z.B. bei k-Means), sondern ergibt sich durch die Ausführung des Algorithmus.
- Es muss nur ein Parameter festgelegt werden (Epsilon).
- Stabilität: Die Ergebnisse sind bei gleichen Ausgangsdaten immer gleich, da der Algorithmus deterministisch vorgeht.

Das Ergebnis des Clusterings hängt maßgeblich von der verwendeten Distanz-Funktion ab. Hierbei ist es schwierig, die Qualität des Clusterings zu beurteilen, da es bei dieser explorativen Cluster-Analyse kein richtig oder falsch, sondern nur ein „nützlich“ bzw. „weniger nützlich“ gibt. Auch die „richtige“ Anzahl der Cluster ist im Vorhinein nicht klar. Wenn jedoch kein (alle Punkte werden als Noise

verworfen) oder nur ein Cluster (alle Punkte im gleichen Cluster) gefunden wird, bietet das Clustering trivialerweise keinerlei Informationsgehalt.

Daher haben wir uns eine Basis-Distanzfunktion erstellt, die aus dem gewichteten Mittel der geografischen Distanz und der Temperatur-Distanz von zwei Messstationen besteht. Durch Verwenden verschiedener Temperatur-Distanzen und Gewichtungen können so mit dem gleichen Basis-Algorithmus unterschiedliche Clusterings mit unterschiedlichen Analyseansätzen ausprobiert werden.

Nach Festlegen einer Temperatur-Distanzfunktion und Gewichtung muss anschließend noch der Parameter Epsilon für das DBSCAN-Verfahren festgelegt werden. Hierfür führen wir eine Suche über verschiedene mögliche Epsilon-Werte durch und ermitteln dabei jeweils die Anzahl der Cluster und des Noise (nicht zugeordnete Datenpunkte). Die Auswahl des optimalen Epsilon erfolgt dann auf Basis dieser Ausgabe manuell, wobei wir jeweils einen Epsilon-Wert gewählt haben, mit dem wir zwischen 3 und 5 Clustern erhalten und die Anzahl des Noise minimieren.

Den ersten Versuch für das Clustering haben wir ausschließlich mit dem Mittelwert und der Standardabweichung der Temperatur jeder Station durchgeführt. Hierfür existiert im Repository ein Video, welches die Clusterentwicklung über die Jahre zeigt. Es wird deutlich, dass Mittelwert und Standardabweichung nicht ausreichend sind, um die Messstationen sinnvoll zu clustern.

### Clustering mit Korrelation

Da wir uns für Cluster mit ähnlichen Temperaturverläufen interessieren, untersuchen wir zunächst ein Clustering anhand der paarweisen Korrelation zwischen den Temperaturwerten der Stationen. Die Korrelation dient dabei als Maß, wie sehr zwei Stationen gleichzeitig einen – relativ zum Stationsmittelwert – hohen bzw. niedrigen Temperaturwert haben.

Für die Analyse verwenden wir als Temperaturdistanz in unserer flexiblen Basisdistanzfunktion die Korrelations-Distanz ( $1 - \text{Korrelation}$  zwischen den Temperaturzeitreihen  $x$  und  $y$ ). Wir versuchen zunächst, ein Clustering mit Gleichgewichtung von Temperaturdistanz und geographischer Distanz zu erhalten. Die Suche über mögliche Epsilon-Werte ergibt dabei:

```
Eps: 0.20 gives 0 clusters and 67 noise
Eps: 0.30 gives 4 clusters and 36 noise
Eps: 0.40 gives 1 clusters and 7 noise
Eps: 0.50 gives 1 clusters and 0 noise
```

Wir sehen also, dass Epsilon 0.2 gar keine Cluster findet und Epsilon 0.4 alle Stationen (bis auf ein wenig Noise) in ein einzelnes Cluster steckt. Wir verwenden daher Epsilon = 0.3 und erhalten das folgende Clustering (nicht zugeordnete Stationen – Noise – werden Grau dargestellt):

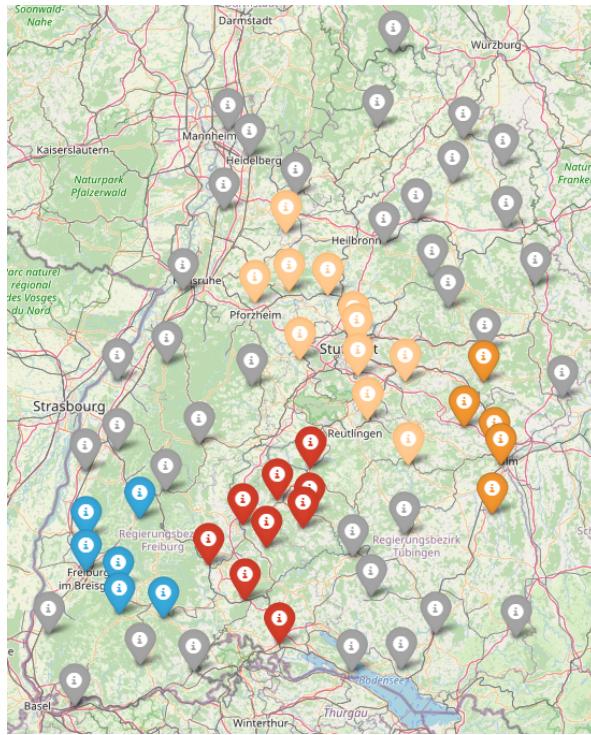
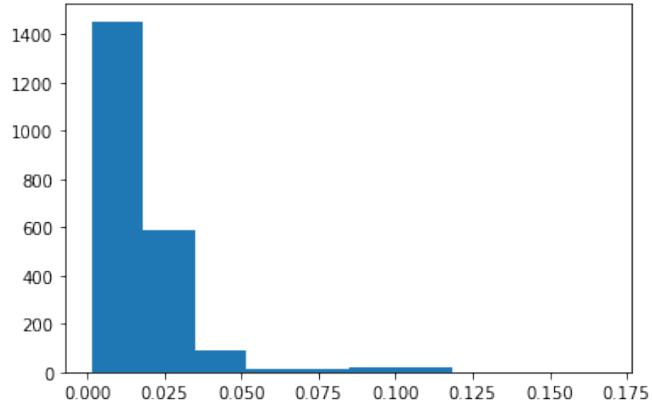


Abbildung 16 Clustering Corr-1 mit Korrelation als Temperaturdistanz und  $w\_dist=1$ ,  $t\_dist=1$ ,  $\text{eps}=0.3$ .

Plotten wir die Verteilung der Temperaturdistanz, so wird deutlich, dass die allermeisten Werte sehr klein sind ( $< 0.05$ ):



Daher liegt es nahe, die Gewichtung der Temperaturdistanz deutlich zu erhöhen. Nach einer manuellen Suche über verschiedene Gewichtungen erhalten wir folgendes Clustering als Ergebnis:

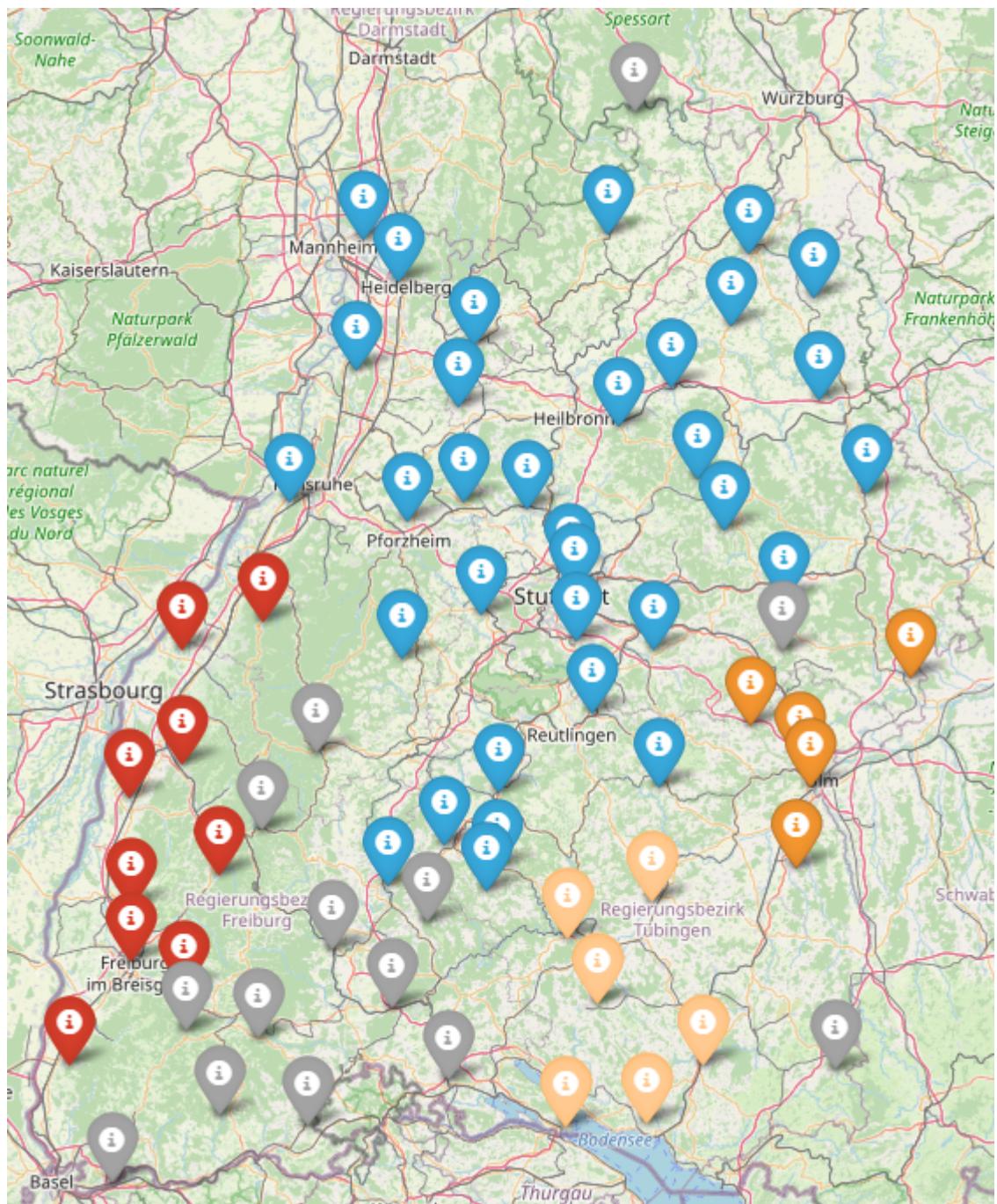


Abbildung 17: Clustering Corr-35 mit Korrelation als Temperaturdistanz und  $w\_dist=1$ ,  $t\_dist=35$ ,  $\text{eps}=0.6$ .

### Clustering mit Cosine-Distance

Als weiteren Ansatz wollen wir nun noch die Abweichung der Stationen vom Landestrend betrachten. Hierzu entfernen wir zunächst den globalen Trend aus den Messreihen der einzelnen Stationen, indem wir den jeweiligen Durchschnittswert an einem Tag über ganz BW von den Messwerten der Stationen abziehen.

Mit diesen Abweichungsreihen berechnen wir nun als Temperaturdistanz jeweils die Cosinus-Distanz. Die Idee ist hierbei, zu berechnen, wie sehr sich die Stationen in ihrer Abweichung von globalem Trend ähnlich verhalten. Um mehr Gewicht auf den zeitlichen Trend als auf einzelne Tagesschwankungen zu legen, kann zusätzlich zuvor ein Resampling auf Monate durchgeführt werden (siehe Jupyter-Notebook).

Auch hier können wieder mit verschiedenen Gewichtungen unterschiedliche Ergebnisse erzielt werden. Ein mögliches Ergebnis ist hier dargestellt:

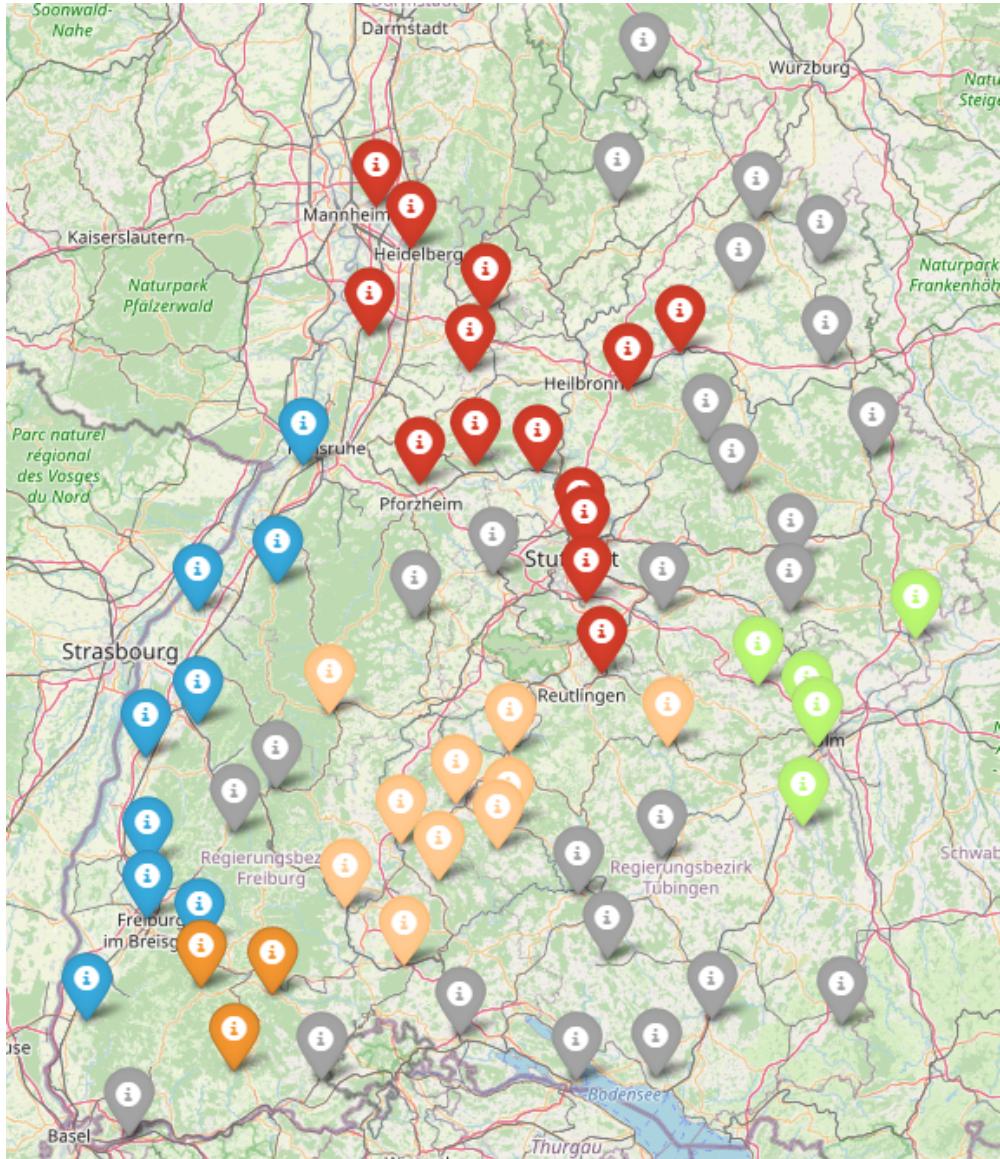


Abbildung 18: Clustering Cosine mit Cosinus-Distanz,  $w\_dist=1$ ,  $w\_tdist=1.5$ ,  $\text{eps}=0.75$ .

## Analyse der relevanten Cluster

### Zeitliche Verläufe der Cluster

Nun wollen wir die gewonnenen Cluster miteinander vergleichen. In diesem Abschnitt gilt folgende Zuordnung von Clusternamen zu den Farben aus den vorherigen Abbildungen:

`0: 'blue', 1: 'red', 2: 'beige', 3: 'orange', 4: 'lightgreen', 5: 'purple', 6: 'lightred', 7: 'black'`

Die Cluster des Clusteringschritts „Cosine“ zeigen im Verlauf der Jahrestemperaturdurchschnitte die Auffälligkeit, dass die Cluster 0 und 1 sowie 2 und 4 sehr ähnlich verlaufen (Abbildung 19). Außerdem weicht der Verlauf des Clusters 3 stark von allen anderen Clustern im Jahr 2019 ab, die Änderung der Temperatur im Cluster 3 ist in diesem Jahr negativ zu allen anderen Clustern korreliert. Grundsätzlich sind die Cluster 3, 2 und 4 sowie 0 und 1 (bis auf das Jahr 2019) um etwa 2°C voneinander getrennt.

Außerdem hat Cluster 0, das geografisch in etwa den Stationen der Oberrheinischen Ebene (vgl. Abbildung 18, blaues Cluster) entspricht, die höchsten durchschnittlichen Temperaturen. Wir erkennen hier, dass das Clusteringverfahren ohne Domänenwissen eine für ihr Klima bekannte Region detektieren konnte.

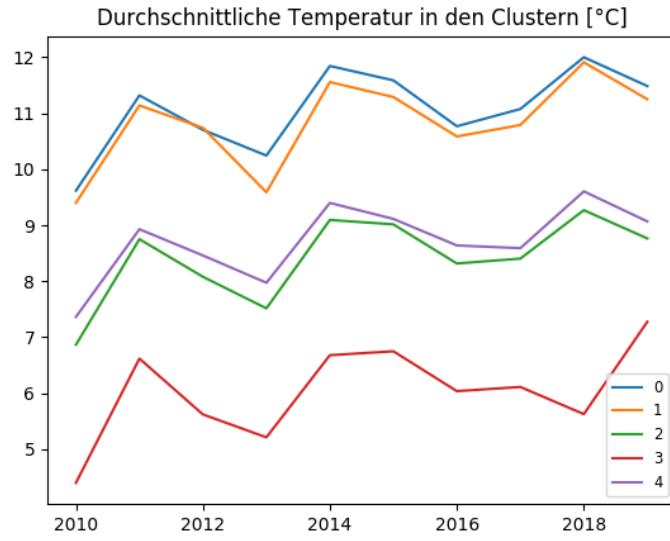


Abbildung 19: Verlauf der jährlichen Temperaturdurchschnitte der Cluster aus dem Clusteringschritt „Cosine“.

Wir setzen ebenfalls statistische Maße auf den Jahresverläufen ein, um weitere Einblicke zu gewinnen (Abbildung 20).

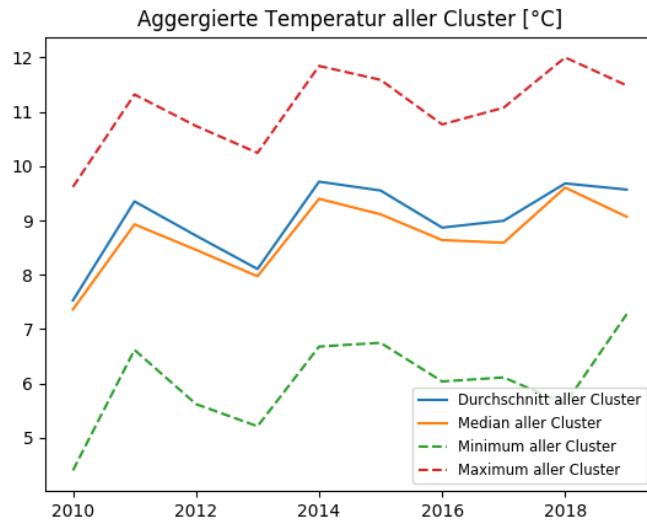


Abbildung 20: Statistische Maße auf dem Verlauf der durchschnittlichen Jahrestemperaturen der Cluster aus dem Clusteringschritt „Cosine“.

Im Plot der Änderungsraten der Jahresdurchschnittstemperaturen lässt sich gut die negative Korrelation von Cluster 3 bezüglich aller anderen Cluster für das Jahr 2019 erkennen (Abbildung 21).

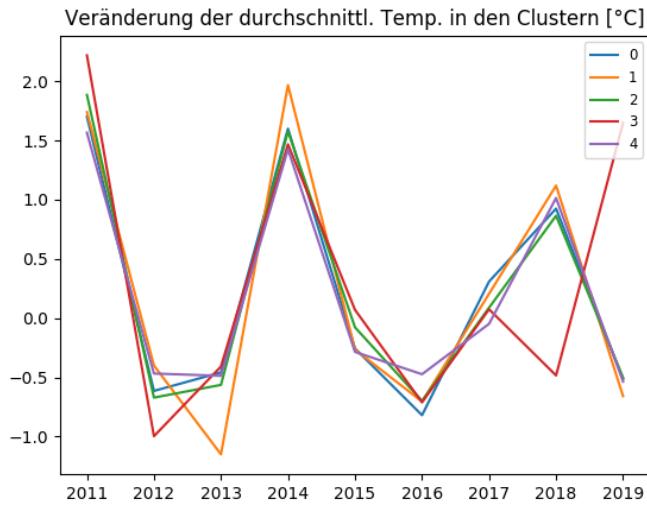


Abbildung 21: Änderungsrate des Verlaufs der jährlichen Temperaturdurchschnitte der Cluster aus dem Clusteringschritt „Cosine“.

Die Cluster aus dem Clusteringschritt „Corr-1“ zweigen ein deutlich anderes Bild im Plot der Jahresdurchschnittsverläufe, als die aus „Cosine“. Von den vier vorliegenden Clustern sind 1 und 3 sehr ähnlich, die anderen beiden haben einen erkennbaren Abstand im Durchschnittstemperaturverlauf. Außerdem sind hier die Verläufe aller Cluster korreliert.

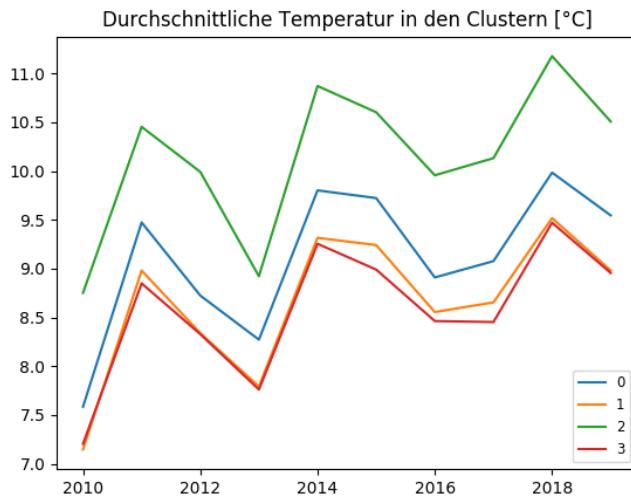


Abbildung 22: Verlauf der jährlichen Temperaturdurchschnitte der Cluster aus dem Clusteringschritt „Corr-1“.

In der weiteren statistischen Analyse der Durchschnittsverläufe der Cluster für „Corr-1“ zeigt sich außerdem, dass der Minimum-Verlauf der Temperaturen, Mittelwert und Median nahe beieinander liegen, während der Verlauf der Maxima nach oben verschoben ist – Cluster 2 ist der Ausreißer, der dies bewirkt, es ist signifikant wärmer als der Mittelwert und weiter nach oben gestreut, als die Ausreißer-Cluster nach unten.

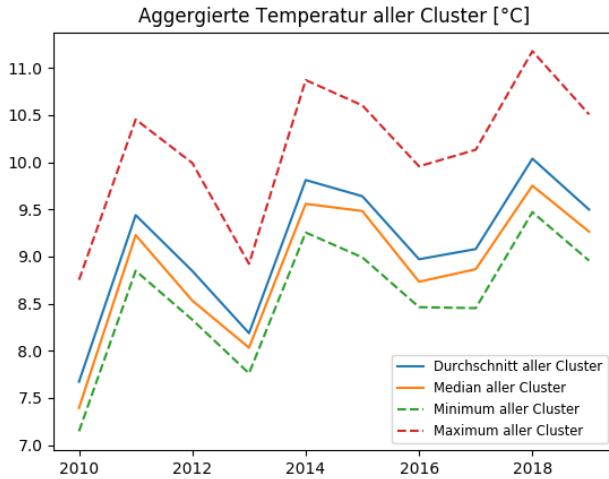


Abbildung 23: Statistische Maße auf dem Verlauf der durchschnittlichen Jahrestemperaturen der Cluster aus dem Clusteringschritt „Corr-1“.

In der Visualisierung der Änderungsrate der Clustertemperaturen lässt sich ablesen, wie bereits zuvor erkannt, dass alle Cluster gut korreliert sind, nur Cluster 2 liegt in den Jahren 2013 und 2014 signifikant oberhalb bzw. unterhalb der anderen Cluster.

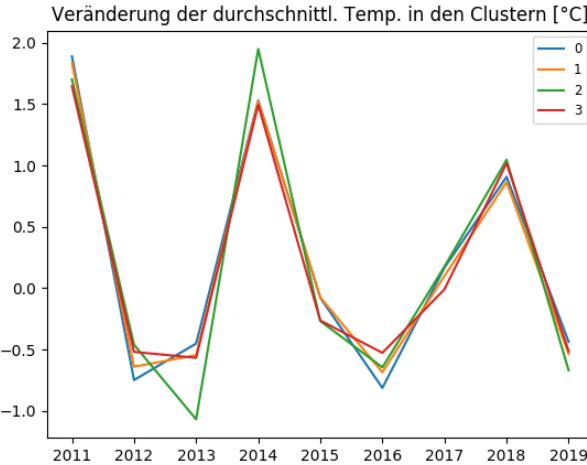


Abbildung 24: Änderungsrate des Verlaufs der jährlichen Temperaturdurchschnitte der Cluster aus dem Clusteringschritt „Corr-1“.

Die Cluster des Clusteringschrittes „Corr-35“ sind im Vergleich zu „Cosine“ und „Corr-1“ relativ gleichmäßig im jährlichen Temperaturverlauf verteilt (Abbildung 25).

Auch „Corr-35“ hat die Stationen in der Oberrheinischen Ebene als eigenes Cluster identifiziert (Cluster 1, vgl. Abbildung 17 rotes Cluster), es weiß stets die höchste durchschnittliche Temperatur unter den Clustern im Jahresverlauf auf.

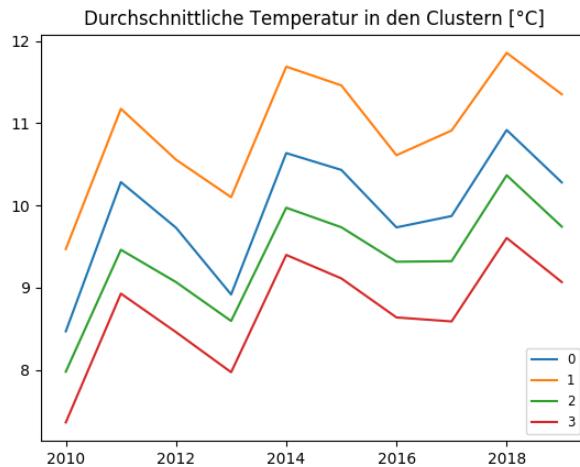


Abbildung 25: Verlauf der jährlichen Temperaturdurchschnitte der Cluster aus dem Clusteringschritt „Corr-35“.

Wie aus dem Plot in Abbildung 25 erwartet, ist die Streuung der Cluster sehr gleichmäßig (vgl. Abbildung 26).

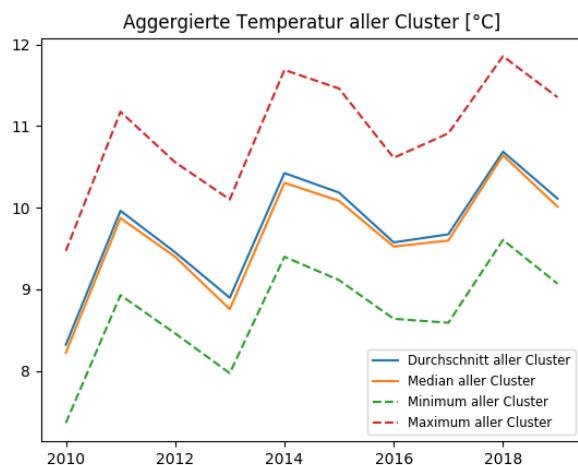


Abbildung 26: Statistische Maße auf dem Verlauf der durchschnittlichen Jahrestemperaturen der Cluster aus dem Clusteringschritt „Corr-35“.

Im Verlauf der Änderungsraten der jährlichen Durchschnittstemperaturen der Cluster zeigt sich außerdem, dass Cluster 0 in manchen Jahren (2013, 2016) ausreißt, jedoch sind die Änderungsraten der Cluster fast immer miteinander korreliert.

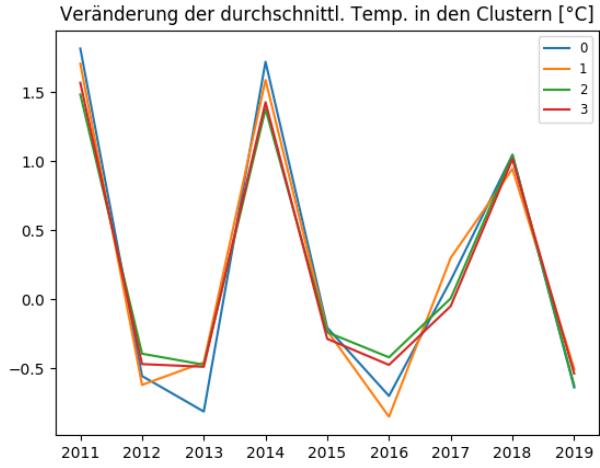


Abbildung 27: Änderungsrate des Verlaufs der jährlichen Temperaturdurchschnitte der Cluster aus dem Clusteringschritt „Corr-35“.

### Geographischen Gesichtspunkte der Cluster

Es fällt auf, dass die örtliche Verteilung der Wetterstationen eine große Rolle für das Clustering spielt. So werden zum Beispiel Wetterstationen im Westen und Osten von BW tendenziell zu dem gleichen Cluster zugeordnet. Abbildung 16 bzw. 17 zeigen zudem, dass das rote bzw. blaue Cluster ausschließlich Wetterstationen beinhaltet, die der oberrheinischen Tiefebene angehören. Ebenfalls können wir alle Wetterstationen in Cluster Orange bzw. Grün dem Mittelgebirge Schwäbische Alb zuordnen. Des Weiteren lässt sich aus dem Cluster Beige in Abbildung 17 erkennen, dass alle dazugehörigen Stationen eine ähnliche Stationshöhe aufweisen (siehe Abbildung 18). Gleiches gilt für Cluster Rot.

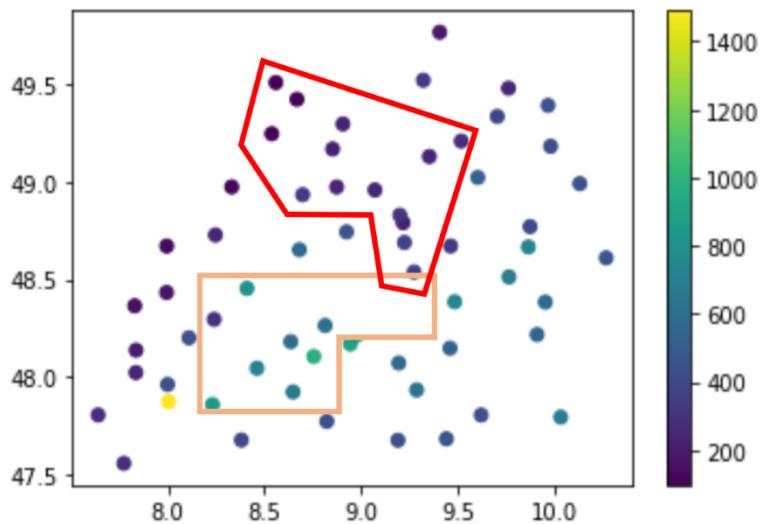


Abbildung 28: Höhe der Wetterstationen

Wir erkennen also, dass unser Clustering Ansatz Cluster erkennt, die sich ebenfalls geographisch in bestimmte Gebiete einteilen lassen. Das ist vor allem interessant, da der gewählte Ansatz ausschließlich die Temperaturverläufe berücksichtigt und nicht beispielsweise die Höhe der Stationen.

## Alternative Clusteringverfahren

Die verschiedenen existierenden Clustering-Verfahren, Distanz- und Ähnlichkeitsmaße und die Verknüpfung der Cluster untereinander sind zahllos und teils komplex. Wir erwägen verschiedene alternative Clusterverfahren, um regionale Cluster zu finden, die sich ähnlich verhalten. Es gibt keine „richtige“ oder „korrekte“ Clusteranalyse, vielmehr werden die Verfahren oft explorativ eingesetzt und mit Fachwissen geeignete Modelle und Zuweisungen „ausprobiert“.

Wir erwägen die folgende Clusterverfahren:

- K-Means: Der K-Means-Clustering-Algorithmus sieht auf den ersten Blick sehr einfach und schnell aus. Aber es gibt einige Herausforderungen, wenn wir es für ähnlich regionale Cluster benutzen. Die Anzahl der Cluster muss im Voraus definiert sein. Die Anzahl ist nicht klar spezifiziert. Dieser Herausforderung begegnen wir mit geschicktem Ausprobieren. Und K-Means hat Probleme Cluster zu erkennen, keine kugelförmige Struktur haben.
- EM (Expectation-maximization algorithm): Eigentlich, der K-Means Algorithmus ist eine vereinfachte Version des EM Algorithmus. Die Anzahl der Cluster für EM Algorithmus muss auch im Voraus definiert sein. EM für dieses Clusteringsmodell verwendet wird ist das Gaussian Mixture Models (GMM) - die Annahme, dass die Punkte des Datensatzes im Allgemeinen der Gaußschen Verteilung folgen.
- Ward: Hierarchische Clusterverfahren werden angewendet, wenn das Ergebnis eine abgestufte Cluster-Struktur aufweisen soll. Typen von hierarchischen Verfahren ist Algorithmus des agglomerierenden hierarchischen Cluster-Verfahrens. Die agglomerativen Clusterverfahren, in denen zunächst jedes Objekt einen Cluster bildet und dann schrittweise die bereits gebildeten Cluster zu immer größeren zusammengefasst werden, bis alle Objekte zu einem Cluster gehören. (Auch bezeichnet als „Bottom-up-Verfahren“). Für die Durchführung einer agglomerativen Clusteranalyse müssen ein Distanz- oder Ähnlichkeitsmaß zur Bestimmung des Abstandes zwischen zwei Objekten und ein Fusionierungsalgorithmus zur Bestimmung des Abstandes zwischen zwei Clustern ausgewählt werden. Dabei ist die Wahl des Fusionierungsalgorithmus oft wichtiger als die des Distanz- oder Ähnlichkeitsmaßes. Ward minimiert die Summe der quadratischen Differenzen innerhalb aller Cluster. Es handelt sich um einen Ansatz zur Minimierung der Varianz, der in diesem Sinne der Zielfunktion k-means ähnelt, jedoch mit einem agglomerativen hierarchischen Ansatz angegangen wird.
- DBScan: DBScan gehört zu Dichtebasierter Verfahren. DBSCAN-Algorithmus überprüft jedes Objekt, ändert seinen Status in "viewed" und klassifiziert es in das Cluster oder Noise, bis schließlich der gesamte Datensatz verarbeitet wird. Die Cluster mit DBSCAN bestimmt können beliebige Formen haben, sind dadurch extrem genau. DBSCAN braucht nicht der Clusteranzahl zu bestimmen. Es wird automatisch ermittelt. DBSCAN hat ein Nachteil. Wenn der Datensatz aus Clustern zu nahe besteht, zeigt die Methode schlechte Ergebnisse. Aber die Verteilung der Stationen in unsere Daten ist jedoch gleichmäßig verteilt. DBSCAN ist ziemlich gut für unsere Daten.

Neben DBScan würde wir hierbei Ward als die alternative Clusterverfahren benutzen. Wenn dies mit den Clustern der Durchschnittsentfernungsmethode verglichen wird, zeigen Stationsgruppierungen im Allgemeinen eine stabilere Struktur. Der Zweck dieses regionalen Clusters ist es, die Temperaturzonen von Deutsch zu bestimmen. Für die Regionalisierung wird eine hierarchische Clusteranalyse ausgewählt. Single-Linkage, Complete-Linkage, Average-Linkage innerhalb eines Clusters und zwischen Clustern und Ward-Techniken werden verwendet, um die für unseren Zweck am besten geeignete Clusteranalysemethode zu bestimmen. Wird-Verfahren ist die wahrscheinlichste akzeptable Ergebnisse in diesem speziellen Fall zu erhalten, wie so oft in klimatologischen Forschung gefunden.

Neben den oben beschriebenen Datenbasierten Ansätzen, können zudem auch semantische Clusteringansätze hinzugezogen werden. Im Speziellen betrachten wir hierbei weitere Informationen über die einzelnen Wetterstationen, wie z.B. die Zugehörigkeit zu einem Ballungszentrum oder die unmittelbare Wassernähe. Wie wir bereits auf Bundeslandebene am Beispiel Hamburg erkennen konnten, kann die Wassernähe durchaus signifikanten Einfluss auf die Temperaturen, vor allem in den verschiedenen Jahreszeiten, nehmen. Deshalb wollen wir dieses Verhalten auch auf Stationsebene in BW untersuchen. Dafür fügen wir zunächst relevante Gewässer (Kanäle, Flüsse, Bäche) zu unserer Datengrundlage hinzu (Abbildung 29) und bestimmen für alle Stationen die kürzeste Distanz zum Wasser.



Abbildung 29: Gewässer Shapes in Deutschland

Ebenfalls können weitere Informationen hinzugezogen werden wie beispielsweise die Bevölkerungsanzahl- und Dichte der jeweiligen Land- bzw. Stadtkreise.

### Bedeutung der Datenqualität und Metainformationen

Anhand der drei Beispiele erkennt man, dass die Daten an teils sehr unterschiedlichen Bereichen erfasst wurden. Vor allem die Stationen Karlsruhe und Berlin-Buch, die beide das jeweilige Stadtclima repräsentieren sollen, weisen sehr unterschiedliche Mikrostandort-Eigenschaften auf. Obwohl beide Wetterstationen mehrere Kilometer vom jeweiligen Stadtzentrum entfernt liegen, kommt es in der direkten Umgebung zu Unterschieden. Die Karlsruher Wetterstation befindet sich mitten in einer

Grünfläche und zusätzlich in unmittelbarer Nähe zu einem See. Die unmittelbare Umgebung der Berliner Wetterstation hingegen ist deutlich stärker bebaut und es herrscht keine Wassernähe vor.

Die Distanz der Wetterstationen zum Stadtzentrum ist ein Faktor, den man bei den Interpretationen der Ergebnisse berücksichtigen muss. Beispielsweise haben unsere Analysen auf S.12 (Abschnitt Ballungszentren) ergeben, dass es keine messbaren Temperaturunterschiede zwischen Wetterstationen in Ballungszentren und Wetterstationen in anderen Gegenden gibt. Bei diesen Ergebnissen müssen wir aber die oben erwähnte Erkenntnis berücksichtigen, dass der mögliche Effekt einer Großstadt auf die Temperaturen nur deshalb nicht von den Wetterstationen gemessen wird, da viele der Wetterstationen außerhalb des Stadtzentrums liegen.

Die Station Großenkneten zeigt eine weitere Besonderheit auf, die bei der Datenanalyse und Interpretation der Ergebnisse beachtet werden sollte. Sie grenzt direkt an eine flächenmäßig sehr große Photovoltaikanlage an. Mehrere Studien untersuchten die Auswirkungen von Photovoltaik Anlagen auf die unmittelbare Umgebung<sup>2</sup>. Sie weisen unter Anderem auf, dass es zu einem sogenannten „Heat Island Effect“ um Photovoltaikanlagen herum kommen kann. Das bedeutet, dass die umliegenden Temperaturen durch die Solarzellen ansteigen können.

Solche lokale Gegebenheiten können unter Umständen die Datenqualität beeinflussen und zu falschen Schlussfolgerungen führen.

Im Hinblick auf die Datenqualität generell muss ebenfalls berücksichtigt werden, dass bei den zugrundeliegenden Daten fehlende Einträge vorliegen und zusätzlich einige Wetterstationen nicht mehr aktiv sind. Beispielsweise kann eine Wetterstation, die nach einem Sommer in einem bestimmten Jahr deaktiviert wurde, dazu führen, dass die Durchschnittstemperatur des jeweiligen Landkreises dadurch positiv verzerrt wird, da die Messdaten für den entsprechenden Winter fehlen.

---

<sup>2</sup>

[https://www.researchgate.net/publication/309121531\\_The\\_Photovoltaic\\_Heat\\_Island\\_Effect\\_Larger\\_solar\\_power\\_plants\\_increase\\_local\\_temperatures\\_Open\\_access\\_httpwwwnaturecomarticlessrep35070](https://www.researchgate.net/publication/309121531_The_Photovoltaic_Heat_Island_Effect_Larger_solar_power_plants_increase_local_temperatures_Open_access_httpwwwnaturecomarticlessrep35070)

## Zusammenfassung und Ausblick

Im Rahmen unserer Arbeit mit den DWD-Lufttemperaturdaten haben wir viele interessante Erkenntnisse gewonnen. So gibt es in Deutschland tatsächlich Regionen, die über bestimmte Zeitspannen betrachtet stets wärmer oder kälter sind als andere. Hierbei spielt die Aufteilung in politische Regionen, über die wir intuitiv sprechen können, jedoch eine große Rolle. Manche Regionen mit auffälligen Temperaturtrends werden auf Bundeslandebene nicht wahrgenommen und umgekehrt fallen andere nur auf, weil sie als alleinstehende politische Region in der Statistik auftauchen. Manche als wichtig erwarteten Faktoren, wie beispielsweise die Zugehörigkeit einer Region zu den Ballungszentren Deutschlands, spielen alleinstehend keine Rolle. Außerdem können Clusteringverfahren ohne Wissen von semantischen Zusammenhängen Regionen der Temperaturentwicklung identifizieren, die auch für Personen mit Domänenwissen sinnvoll erscheinen.

Wetter und Klima sind für den Alltag der Menschen wichtiger als je zuvor und Data-Science-Techniken können wertvolle Einblicke liefern. Regionale Lufttemperatur-Trends sind sehr interessant, aber nur eine Facette der Wetter- und Klimaentwicklung. Weitere Datenreihen wie beispielsweise Luftfeuchtigkeitsverläufe, Niederschlagsmengen, Anzahl von Unwetterereignissen und andere Merkmale könnten eine ganzheitlichere Analyse erlauben.

Über die Data-Science hinaus ist natürlich weitere Forschung und vor allem Entwicklung von Softwarewerkzeugen notwendig, um Domänenexperten Datenanalyse einfacher zugänglich zu machen, wie es sich beispielsweise Cadenza als Ziel setzt.