Data Mining

Versuch 1 Energy Data

von
Armin Schwarz (as219)
Florian Tatzel (ft020)
Marc Walter (mw136)

Das git Repository mit den Quellcodes findet sich unter https://github.com/floriant/DataMining
Alle Dateien zu diesem Versuch befinden sich im Unterordner Versuch_1.

2 Durchführung Teil 1: Energieverbrauch und CO2-Emmission

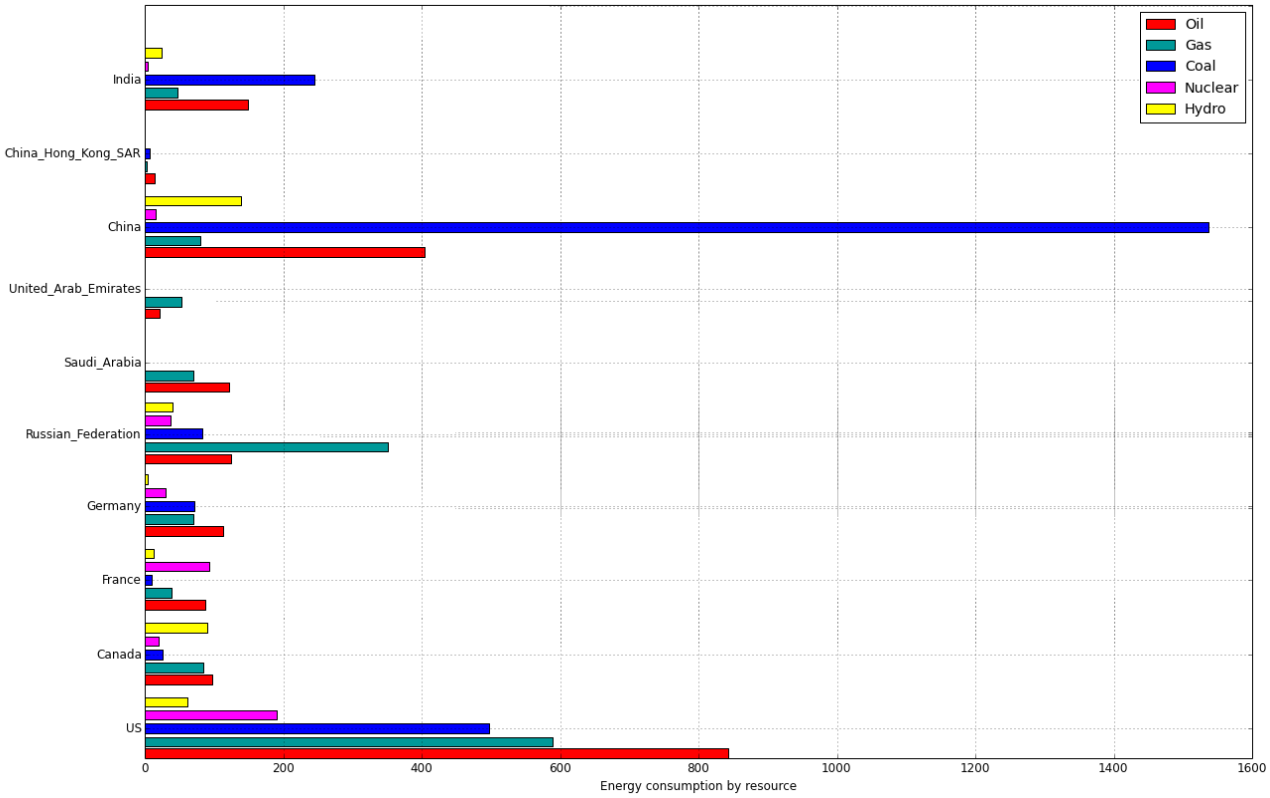
2.1 Datenverwaltung und Statistik

2.1.1 Daten

Ausgehend von der implementierten Visualisierung des Energieverbrauchs der Länder: Nennen Sie die 3 Ihrer Meinung nach interessantesten Beobachtungen.

1. China verbraucht ungefähr so viel Kohle wie die restlichen Länder der Statistik zusammen
2. Es gibt Länder, die manche Ressourcen überhaupt nicht verwenden (Nuclear und Hydro bei den Vereinigten Arabischen Emirate oder Saudi Arabien
3. Wenn man China und die USA aus der Grafik heraus lässt, kann man die Werte gut lesen, da die meisten anderen Länder vergleichbar viel Energie verbrauchen.

Auszug aus der graphischen Darstellung des Energieverbrauchs



Die komplette graphische Darstellung findet sich im Ordner **Versuch_1\doc** in den Dateien **EnergyMix.png** und **EnergyMix.pdf**.

Zu dieser Aufgabe gehört die Datei **Versuch_1\Program\appendGeoCordinates.py**.

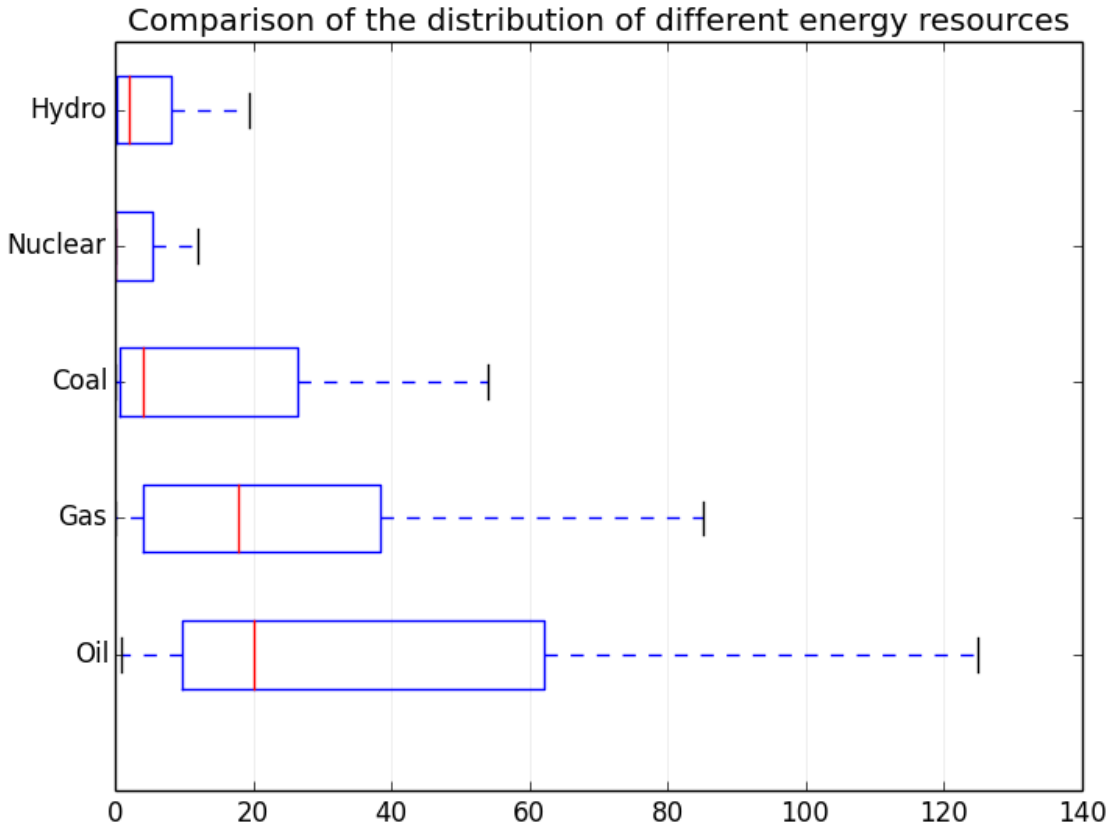
2.1.3 Statistik der Daten

1. Erklären Sie sämtliche Elemente eines Boxplot (allgemein).

Ein Boxplot ist ein Diagramm, das gut die Streuung verschiedener Daten anzeigen kann.
In der (bei uns blauen) Box liegen die mittleren 50% der Daten, genauer gesagt zwischen dem unteren Quartil (25% der Daten) und dem oberen Quartil.
Der (bei uns rote) Strich innerhalb der Box kennzeichnet den Median, also den Wert der in der Mitte steht, wenn alle Werte aufsteigend sortiert werden.
Die Whisker an den beiden Seiten kennzeichnen weitere Werte, die unterhalb/überhalb des Quartils liegen, aber in unserem Fall noch im Wertebereich von maximal 1,5 Mal der Länge des jeweiligen Quartils liegen.
Außerhalb der Whisker existieren noch Ausreißer (fliers), die für die allgemeine Statistik meist keine Relevanz haben. Deswegen sind sie in diesen Boxplots auch nicht dargestellt, allerdings wurden auch alle Schaubilder auch mit Ausreißern im Ordner **Versuch_1\doc** erzeugt.

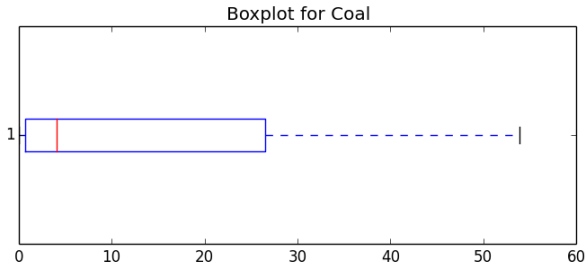
2. Diskutieren Sie die im Boxplot angezeigte Statistik der Energieverbrauchdaten.

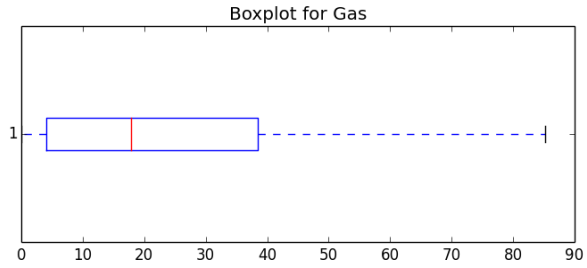
In der Übersicht lässt sich erkennen, dass es viele Länder gibt, die relativ viel Öl verbrauchen, aber der Median von Gas und Öl weltweit in einem vergleichbaren Rahmen liegt.

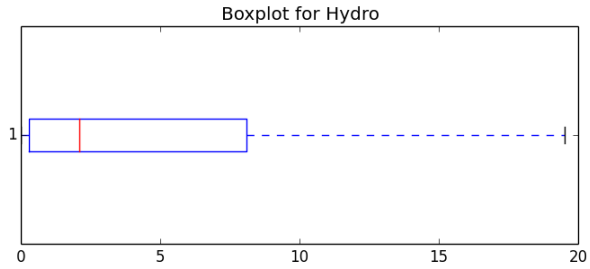


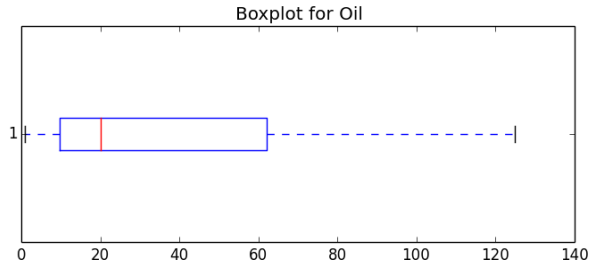
Boxplots für die einzelnen Energietypen

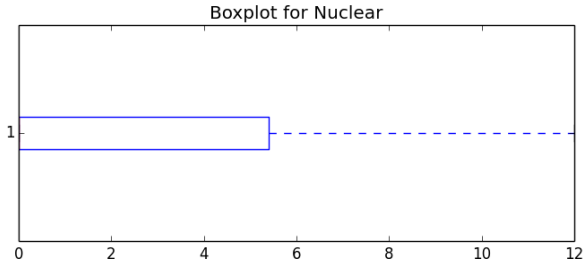
In diesen Plots sind die Ausreißer (fliers) für bessere Lesbarkeit der Grafiken nicht mit ausgegeben. Die Grafiken wurden auch mit Ausreißern erzeugt und befinden sich im Ordner **Versuch_1\doc**.











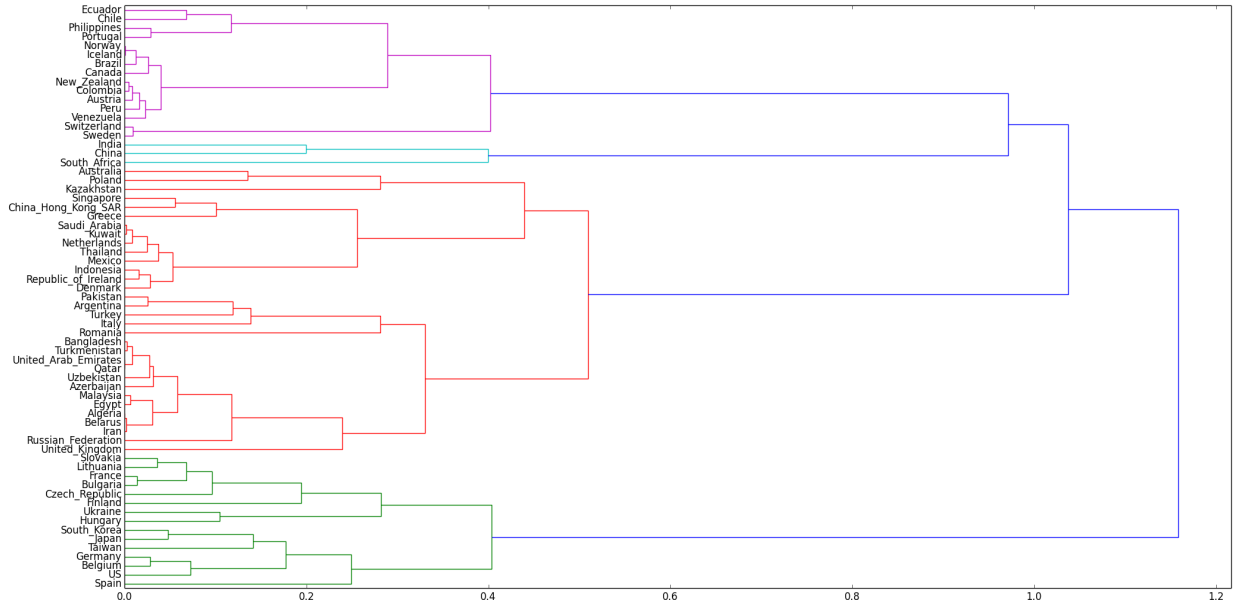
Quellcode zu dieser Aufgabe befindet sich in der Datei **Versuch_1\Program\energyStatistics.py**.

2.2 Anwendung von Verfahren des unüberwachten Lernens auf Energieverbrauchsdaten

2.2.1 Hierarchisches Clustering

Dendrogram

In diesem Schaubild lässt sich ablesen, welche Werte zueinander die geringste Distanz haben, und sich somit am ähnlichsten sind.



Die Gesamtgrafik ist unter **Versuch_1\doc\dendrogramm.png** gespeichert.

1. Was wird beim Standardisieren gemacht? Welcher Effekt könnte ohne Standardisieren beim Clustering eintreten (insbesondere wenn die euklidische Metrik verwendet wird)?

Beim Clustering werden die Eigenschaften von Objekten gemessen. Da die Eigenschaften aber unterschiedlich skalieren können, müssen sie standardisiert werden. Die verwendeten Eigenschaften werden daher mit einer Distanz-Funktion auf ein ähnliches, und damit untereinander vergleichbares, Maß gebracht.

Eine Fehlende Standardisierung würde zu einem verfälschten Ergebnis führen, da der Einfluss von Kohle auf die Emmision wesentlich höher ist, als beispielsweise der von Nuklear-Energie. Die euklidische Metrik würde die Distanz zwischen den Eigenschaften angeben und somit hätten sehr hohe Werte eine unverhältnismäßig gößere Gewichtung.

2. Erklären Sie die beim hierarchischen Clustering einstellbaren Parameter linkage-method und metric. Welche Metrik ist Ihrer Meinung nach für diese Anwendung geeignet? Warum?

Der Parameter “metric” gibt die zu verwendende Metrik für die Distanzfunktion an. Als Default ist hier die Metrik “euclidean” angegeben. Diese kann bei starken Wertschwankungen der Eigenschaften aber zu Problemen (siehe Frage 1) führen.

Daher wird die Metrik “correlation”, verwendet. Diese Metrik erlaubt es, eine Distanzberechnung von stark skalierenden Eigenschaften vorzunehmen.

Die linkage-method ist hingegen ein Wert, welcher die Distanzen von Clustern untereinander berechnet. Dabei wird in diesem Versuch die Methode “average” verwendet, da sie den Abstand zu den Cluster-Mittelpunkten berechnet.

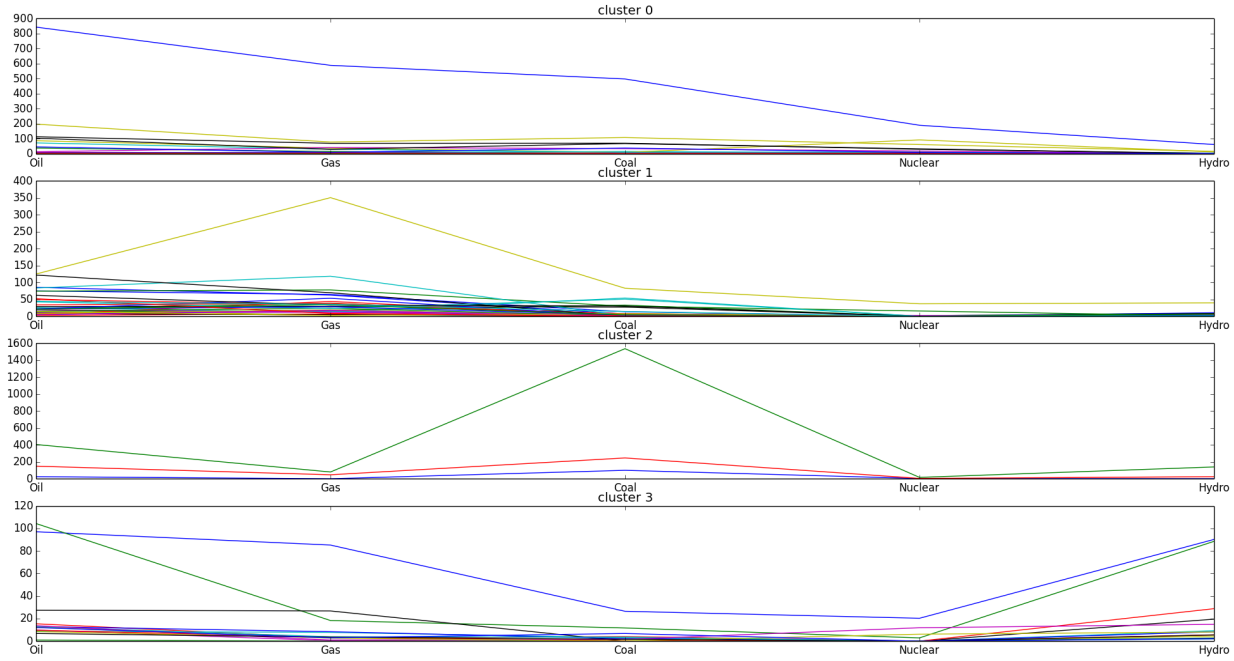
3. Welches Land ist bei der Energieverteiung Deutschland am ähnlichsten?

Am ähnlichsten zu Deutschland ist das Land Belgien.

4. Charakterisieren Sie die 4 Cluster. Was ist typisch für die jeweiligen Cluster?

- Cluster 0: Länder haben eine ausgewogene Enegie-Ressourcenverteilung.
- Cluster 1: Diese Länder verwenden vorwiegend Gas als Enegiequelle.
- Cluster 2: Diese Länder verwenden vorwiegend Kohle als Energiequelle.
- Cluster 3: Diese Länder verwenden die Klassischen Energieformen (Öl, Gas, Kohle) und zusätzlich verstärkt Wasserkraft.

Diese Grafik veranschaulicht die Gewichtung der einzelnen Eigenschaften für jedes Cluster.

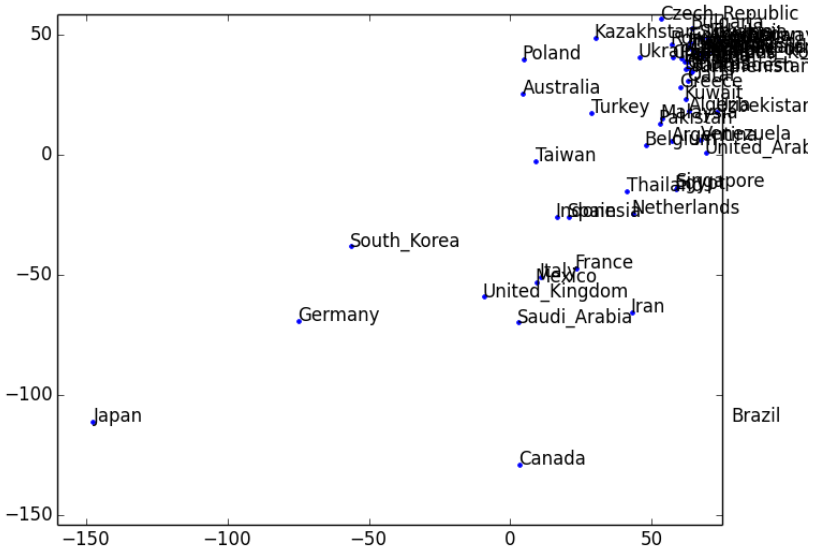


Die Gesamtgrafik ist unter **Versuch_1\doc\individualClusters.png** gespeichert.

Der Code dieser Aufgabe ist in der Datei **Versuch_1\Program\energyFeatureSelection.py**.

2.2.2 Dimensionalitätsreduktion

Die Statistiken wurden auf zwei Dimensionen reduziert, um sie in einem 2D-Plot darstellen zu können. Hier ein Ausschnitt dieses Plots.



Die Gesamtgrafik ist unter **Versuch_1\doc\screenshot_EnergyMix.png** gespeichert.

1. Welches Land ist nach dieser Darstellung Deutschland am ähnlichsten?

Am ähnlichsten scheint Südkorea zu sein, da die euklidische Distanz am geringsten ist.

2. Warum entspricht die hier dargestellte Ähnlichkeit nicht der im oben erzeugten Dendrogram?

Generell kann es zu Veränderungen kommen, da die Anzahl der Dimensionen nicht ohne Informationen zu verlieren reduziert werden können.
Bei dieser Ansicht spielt auch der Gesamtenergieverbrauch eine große Rolle, welcher im Dendrogram nicht berücksichtigt wird. Zudem werden die Verhältnisse der Energiemixe nicht gewichtet, sondern die reinen Zahlenwerte der Energien.

Der Quellcode befindet sich in der Datei **Versuch_1\Program\energyReduceDim.py**

2.3 Überwachtes Lernen: Schätzung der CO2-Emmission

2.3.1 Feature Selection

1. Welche 3 Merkmale haben den stärksten Einfluss auf das Ausgabemerkmal CO2-Emmission? Wie groß sind die vom Programm ausgegebenen Scores?

Den höchsten Einfluss auf das Ausgabemerkmal haben hierbei Oil, Coal und Hydro.

Die vom Programm ausgegeben Scores betragen:

Oil	Gas	Coal	Nuclear	Hydro
220.01015147	46.00222955	378.26688078	34.57208601	79.04540111

Der Quellcode zu dieser Aufgabe befindet sich in der Datei
Versuch_1\Program\energyFeatureSelection.py

2.3.2 Regression mit Epsilon-SVR

1. Optimieren Sie die SVR-Parameter C und Epsilon so dass der Score in der Kreuzvalidierung minimal wird. Welche Werte für C und Epsilon liefern das beste Ergebnis?

Für die Optimierung der Werte C und Epsilon wurde kein automatisches Verfahren gefunden. Daher wurden Wertbereiche untersucht und mit ihnen der minimale Score ermittelt.
Dabei wurde für C der Wertbereich von 1-7 in 0.1 Schritten untersucht, bei Epsilon lag der Bereich bei 0.1-0.2 in 0.001 Schritten.
Der Minimale Score von **0.0303** wurde mit **C=6.7** und **Epsilon=0.133** ermittelt.

2. Für das SVR-Objekt können die Koeffzienten der linearen Abbildung, welche durch die trainierte SVR realisiert wird, ausgegeben werden: meineSVR.coef_ . Notieren Sie diese Koeffizienten für die beste SVR.

Die Koeffizienten lauten wie folgt:

```
Oil: -3.06911037
Gas: -2.34782548
Coal: -3.96083151
Nuclear: -0.000772739642
Hydro: -0.00120585976
```

3. Welchen Aufschluss geben diese Koeffizienten über den Einfluss der einzelnen Eingangsmerkmale auf das Ausgangsmerkmal?

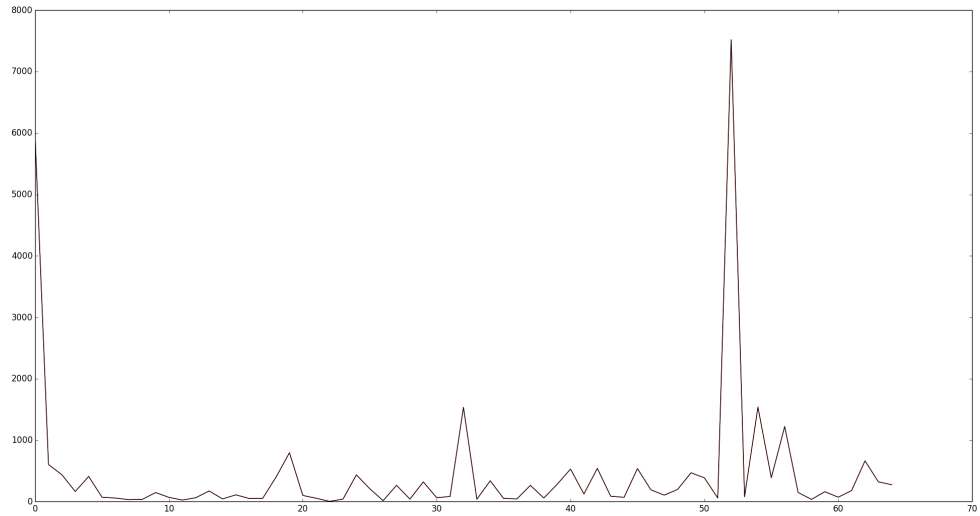
Die Koeffizienten geben Aufschluss über die Gewichtung eines Eingabemerkmals auf das Ausgabemerkmal an. Anhand der zurückgelieferten Daten lässt sich feststellen, dass die Merkmale Oil, Gas und Coal gegenüber Nuclear und Hydro ein wesentlich höheren Einfluss auf

die CO2 Emmisionen hat. Dabei hat Coal den höchsten Einfluss und Nuclear den niedrigsten Einfluss.

4. Wie groß ist die mittlere absolute Differenz zwischen Soll- und Ist-Ausgabe für die beste SVR? Diskutieren Sie dieses Ergebnis.

Die mittlere absolute Differenz (MAD) beträgt **0.130**.
Dieser Wert weist auf eine geringe Abweichung der Soll-Ausgabe von der Ist-Ausgabe hin.
Dies spiegelt sich auch im Plot der beiden Ausgaben wieder, und weist auf eine gut angepasste Vorhersage hin.

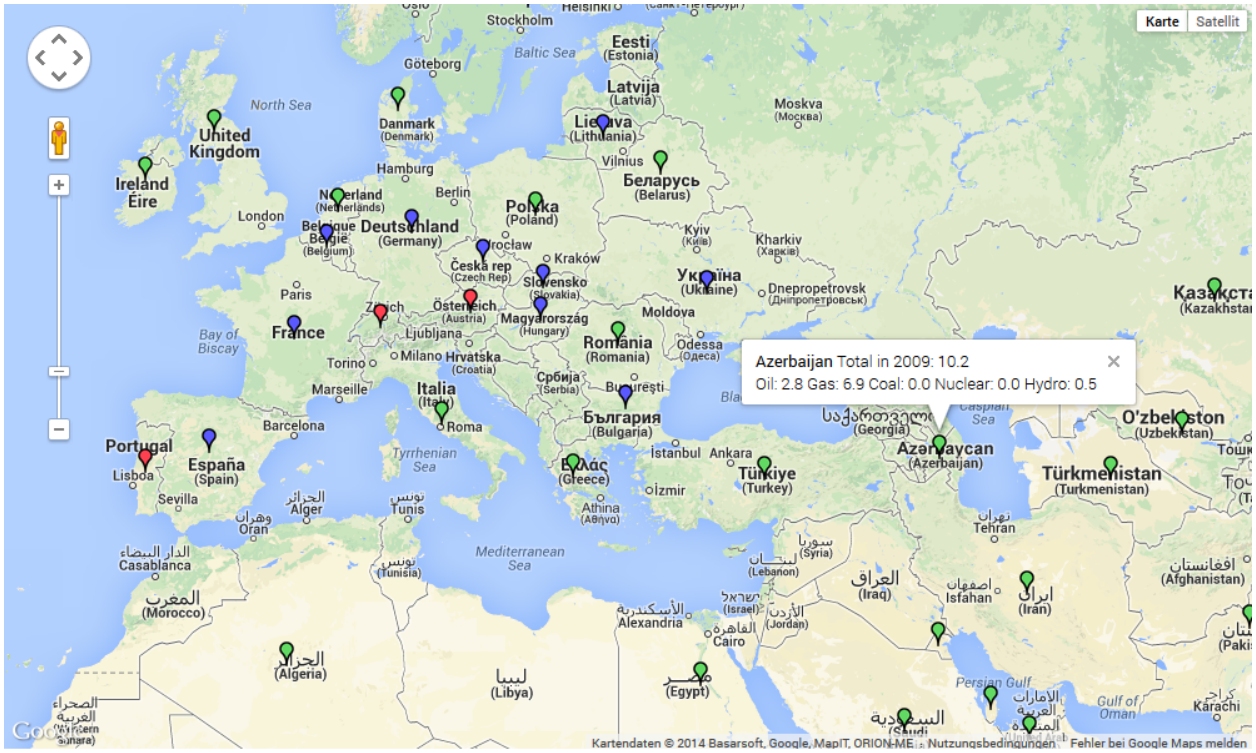
Der Plot zeigt beide Graphen, wobei nur eine Linie sichtbar ist, aufgrund der Geringen Abweichung.



Die Gesamtgrafik ist unter **Versuch_1\doc\energyPrediction.png** gespeichert.

Der Quellcode zu dieser Aufgabe befindet sich in Datei
Versuch_1\Program\energyPrediction.py

2.4 Visualisierung des Clusterings in Google Maps



Diese Grafik zeigt einen Ausschnitt des in der Datei **Versuch_1\doc\EnergyMix2009.htm** visualisierten Clusterings.

Zur Erzeugung werden aus dem Ordner **Versuch_1\Program** die beiden Dateien **clusters2GoogleMaps.py** und **pymaps.py** benötigt.

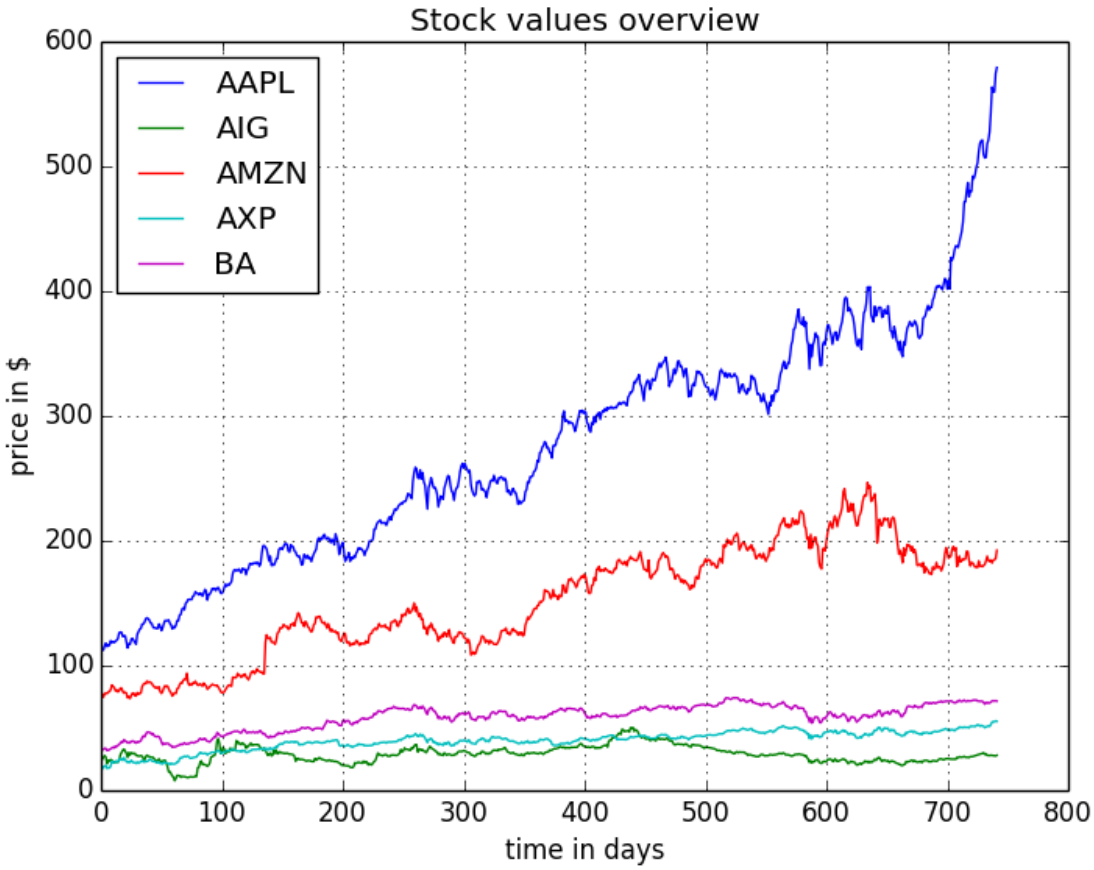
3 Durchführung Teil 2: Vorhersage und Clustering auf Finanzdaten

3.1 Zeitreihenschätzung: Vorhersage des Aktienkurses

3.1.1 Datenbeschaffung

Die Datei **Versuch_1\Program\b101_stockMarketFile.py** enthält denCode für das Beschaffen der Aktienkurse und das Speichern in der Datei **Versuch_1\res\effectiveRates.csv**.

3.1.2 Kursvorhersage mit SVR



Die Aktienkurse einiger Firmen finden sich über den gesamten Zeitraum als Plot in der Datei **Versuch_1\doc\stock_overview.png**.

1. Wie müssen die Datenvektoren des Vorhersagezeitraums aufgebaut werden?

Die bestehenden Datenvektoren des Vorhersagezeitraums müssen so ergänzt werden, dass jeder vorhergesagte Wert einen eventuell bestehenden Wert überschreibt und die weiteren Vorhersagen auf diesem aufbauen.

2. Für welche Werte von Time Delay, SVR-Parameter C und SVR-Parameter epsilon erreichen Sie die beste Vorhersage? Wie groß ist in diesem Fall der MAE?

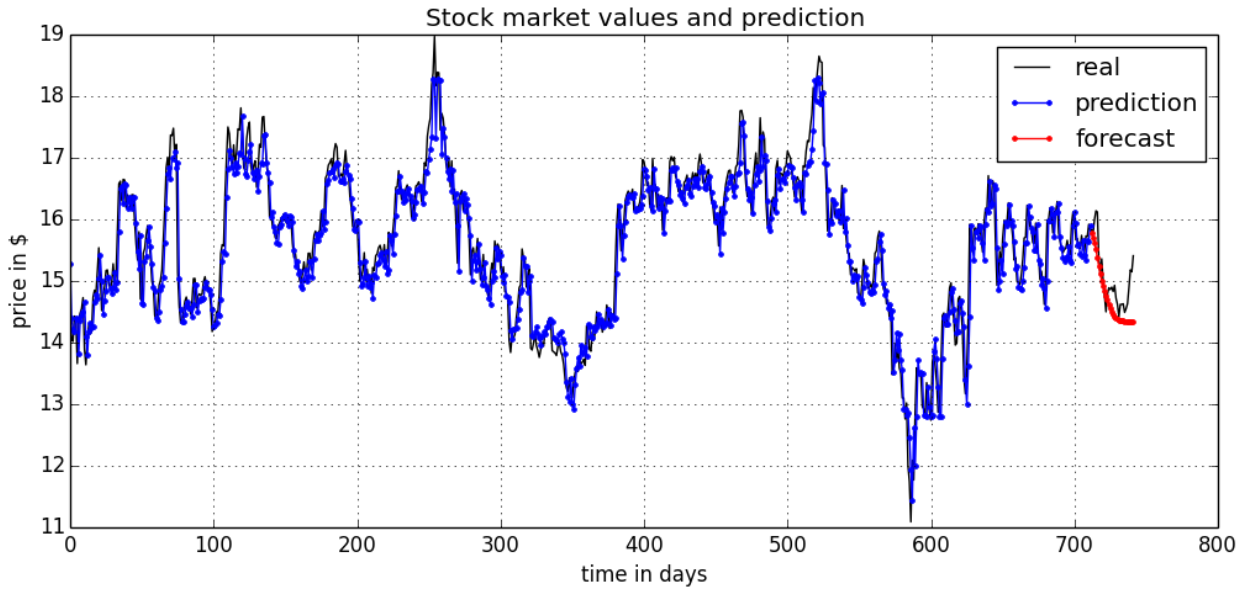
Getestet wurden Werte für time_delay zwischen 1 und 35, c zwischen 400 und 500 und epsilon zwischen 0.1 und 0.9 mit einer Schrittweite von 0.05.

Die besten Ergebnisse liefern die Parameter

```
time_delay = 1
c = 400.0
epsilon = 0.85
```

und errechnen einen MAE von **~0.335**.

Da die Vorhersage generell besser wird, wenn weniger Werte für die Berechnung verwendet werden, kann der Versuch die Aktienkurse mit den erhaltenenen Daten richtig vorherzusagen als gescheitert betrachtet werden.



Diese Grafik ist in der Datei **Versuch_1\doc\stockpredict.png** gespeichert.

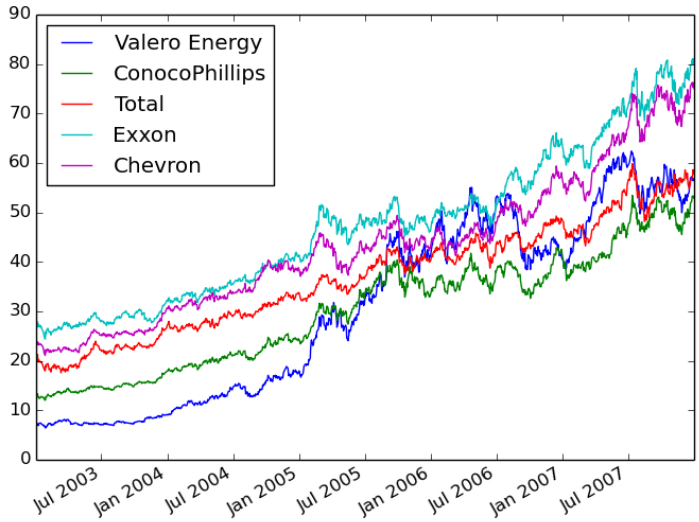
Quellcode befindet sich in der Datei **Versuch_1\Program\b102_stockMarketPrediction.py**.

3.2 Clustering der Aktienkursverläufe

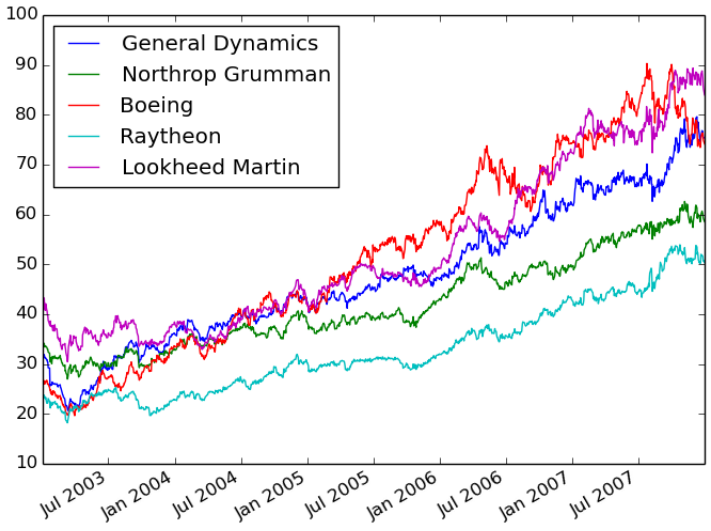
1. Analysieren Sie die Clusterzuweisungen. Gehören die Unternehmen, welche in einem Cluster zusammenfallen, irgendwie zusammen?

Die Cluster repräsentieren grob Branchen, da sich die Kurse der Unternehmen innerhalb einer Branche teilweise sehr ähnlich entwickeln.

Energie (Cluster 4):



Rüstung (Cluster 10):



Die Grafiken sind in den Dateien **Versuch_1\doc\stockMarketCluster%d.png, %0-10** gespeichert.

In der Datei **Versuch_1\Program\b103_stockMarketClustering.py** befindet sich der Quellcode für diese Aufgabe.