# Dataminig - Versuch 1

Lukas Spieß, Mathis Hoffmann

11. November 2013

# 2 Durchführung Teil 1: Energieverbrauch und CO2-Emmission

## 2.1 Datenverwaltung und Statistik

### 2.1.2 Einlesen der Daten, Hinzufügen der GPS Koordinaten, Abspeichern in neuer Datei

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

   **Hinweis:** Um Extreme besser auslesen zu können wurden vereinfachte Graphen mit den jeweils 7 Peaks für jede Energieart zusätzlich angefügt.

   - Einzelne, sehr starke Ausreisser nach oben, die eine Visualisierung schwierig machen, da sie die Skala in y-Richtung sehr stark dehnen.
   - China erzeugt weltweit am meisten Energie aus Kohle, mehr als dreimal so viel, als der zweite in der Rangfolge, die USA. Gleichzeitig hat China auch die größte Erzeugung von Energie aus Wasserkraft.
   - Frankreich sticht ausnahmsweise bei der Energieerzeugung aus Kernenergie hervor, liegt hier aber dennoch deutlich hinter den USA.

### 2.1.3 Statistik der Daten

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

   Ein Boxplot besteht in jedem Fall aus einem Rechteck (der Box) und zwei Linien (die „Antennen“). Dabei ist die Größe der Box so gewählt, dass sie genau über die mittleren 50% der Daten ausdehnt. Im Gegensatz zur einheitlichen Definition der Box gibt es für die Länge der Antennen unterschiedliche Definitionen. Standardmäßig beträgt die Länge der Antennen in *Matplotlib* (das wir in unserem Beispiel indirekt über dessen Integration in *Pandas* verwendet haben) das 1,5-fache des Interquartilsabstands[1]. Allerdings enden die Antennen nicht genau an diesem Punkt sondern an dem letzten Datenwert, der noch innerhalb dieses Bereiches liegt. Zusätzlich wird der Median der Daten noch also Linie in die Box

[1] Der Interquartilsabstand ist die Differenz der Quartile $Q_{75}$ und $Q_{25}$, beschreibt also wiederum genau den Bereich in dem 50% der Daten liegen.

eingezeichnet. Ausreißer, also solche Daten die Außerhalb der Antennen liegen, werden zusätzlich eingetragen.

Wir haben uns dafür entschieden, alle Energieformen in einem *Plot* (mit entsprechenden *Subplots*)darzustellen, um eine bessere Vergleichbarkeit der einzelnen Energieformen zu erreichen.

2. *Diskutieren Sie die im Boxplot angezeigte Statistik der Energieverbrauchsdaten.*

   **Hinweis:** In der beigefügten Grafik[2] haben wir die Darstellung etwas skaliert, um eine bessere Lesbarkeit zu erreichen. Dadurch ist ein Ausreißer der Energieform Öl (bei knapp 900) nicht dargestellt.

   Die Boxplots lassen u.A. folgende Schlussfolgerungen zu:

   - Die am meisten verbreiteten Energieformen sind Öl und Gas.
   - Die Streuung im Verbrauch der einzelnen Länder ist sehr hoch. Das lässt sich aus den langen Antennen (insbes. nach oben) ablesen. Darüber hinaus gibt es sehr extreme Ausreißer.
   - Atomenergie und Kohle kommen ist jeweils der durchschnittliche Verbrauch relativ niedrig. Dafür gibt es jedoch besonders viele Ausreißer. Das deutet darauf hin, dass Länder entweder sehr stark auf diese Energieformen setzen, oder aber nahezu gar nicht auf sie zurück greifen.

## 2.2 Anwendung von Verfahren des unüberwachten Lernens auf Energiever- brauchsdaten

### 2.2.1 Hierarchisches Clustering

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

   Ohne eine Standardisierung würden unterschiedlich skalierte Attribute verschieden stark gewichtet, was zu einer Ergebnisverzerrung führen könnte. In unserem Beispiel würde bspw. der Energieverbrauch aus durch Öl gewonnener Energie einen viel stärken Einfluss auf das Clustering haben, als das der Energieverbrauch, der durch Wasserkraft abgedeckt wird.

   Dieser Effekt würde insbesondere beim Einsatz der euklid'schen Metrik auftreten, da diese, im Gegensatz zu bspw. der Pearson-Metrik, ein Distanzmaß ist und somit die absoluten Distanzen zwischen den einzelnen Clustermittelpunkten von entscheidender Bedeutung sind.

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 *linkage-method* gibt die zu verwendende Methode zur Berechnung des Abstandes zweier Cluster vor. Es stehen im wesentlichen *single* (bei der einfach der Abstand der einander am nächsten liegenden

[2] Vgl. Grafik *2.1_Boxplots.png*

Punkte der beiden Cluster verwendet wird), *complete* (ähnlich *single* - nur, dass die am weitesten voneinander entfernt liegenden Punkte maßgebend sind) und *average* (hier wird die mittlere Distanz aller Clusterpunkte des einen Clusters zu allen des jeweils anderen Clusters berechnet) zur Verfügung.

Der Parameter *metric* gibt die Methode, mit der der Abstand zweier Datensätze zueinander berechnet wird, vor. In unserem Fall macht die Verwendung der Metrik *correlation* Sinn, weil wir skalare Datensätze clustern und somit der Abstand der einzelnen Datensätze mit dieser Metrik direkt berechenbar ist.

3. *Welches Land ist bezüglich des Verbrauchs der hier betrachteten Energiequellen Deutschland am ähnlichsten, wenn für die linkage-method average und die Metrik correlation konguriert wird?*

   Belgien

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

   **Cluster 1:** hoher Anteil an Energie aus Öl, Kohle und Atomkraft; kaum Wasserkraft

   **Cluster 2:** sehr hoher Anteil an Energie aus Gas; kaum Wasserkraft

   **Cluster 3:** sehr hoher Anteil an Energie aus Kohle; kaum Wasserkraft; keine Atomkraft

   **Cluster 4:** mäßiger Anteil an Energie aus Öl und Gas; wenig Kohle und Gas; signifikant hoher Anteil aus Wasserkraft

### 2.2.2 Dimensionalitätsreduktion

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

   Südkorea

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

   Da in 2.2.1 3.) von den Isomap Standardeinstellungen abgewichen wird und als Metrik stattdessen *correlation* verwendet wird.

## 2.3 Überwachtes Lernen: Schätzung der CO2-Emmission

### 2.3.1 Feature Selection

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

   *entfällt*

#### 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?*

   Es wurden im Versuch für C und Epsilon alle Werte zwischen 0,1 und 2,0 (Schrittgröße: 0,1), beziehungsweise 0,01 und 0,2 (Schrittgröße 0,01) ausprobiert. Dabei haben sich als optimale Werte C=1,9 und Epsilon=0,01 ergeben.

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

   Die Koeffizienten sind -3.06917079e+00, -2.34851788e+00, -3.96082193e+00, 6.28903777e-04 und 6.77046323e-04

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

   Je größer C ist, desto stärker wird für eine Abweichung, die Außerhalb der Epsilonumgebung liegt bestraft. Der Parameter Epsilon gibt die maximale Entfernung (oder Toleranz) vom tatsächlichen Wert an, bei der eine Bestrafung für die Abweichung der Vorhersage von diesem Wert entfällt

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 ist mit 0.121 ein überraschend geringer Wert. Auf Grund dieses Wertes lässt sich schließen, dass es sich hier um ein sehr präzises Vorhersagemodell handelt.

# 3 Durchführung Teil 2: Vorhersage und Clustering auf Finanzdaten

## 3.1 Zeitreihenschätzung: Vorhersage des Aktienkurses

### 3.1.2 Kursvorhersage mit SVR

**Hinweis zur Aufgabendurchführung:** Aufgrund der Tatsache, dass diese Aufgabe nicht obligatorisch ist, haben wir sie, rein aus Interesse, noch recht gegen Ende der verfügbaren Zeit (zu Hause) durchgeführt und konnten daher nicht mehr auf externe Hilfe zurückgreifen. Aufgrund der knappen Zeit konnten wir so auch nicht mehr die empfohlene Rücksprache mit dem Betreuer nach Anlegen der Trainingsdaten halten. Dennoch möchten wir im folgenden unseren Lösungsweg dieser Aufgabe vorstellen, der jedoch keinen Anspruch auf Korrektheit hat.

1. *Überlegen Sie sich genau, wie die Datenvektoren des Vorhersagezeitraums (also die Vektoren, die der Methode prediction() des trainierten SVR-Modells übergeben werden), aufgebaut werden müssen.*

   vgl. Sourcecode

2. *Für welche Werte von Time Delay, SVR-Parameter C, SVR-Parameter epsilon erreichen Sie die beste Vorhersage? Wie groß ist in diesem Fall der MAE? Erzeugen Sie für diese optimierten Werte den Plot des tatsächlichen und des geschätzten Kursverlaufs und speichern Sie diese Grak unter dem Namen stockpredict.png.*

   Um diese Frage beantworten zu können haben wir ein extra Programm[3] geschrieben, das eine gewisse Menge an Möglichkeiten nach dem *brute-force-Ansatz* ausprobiert und am Ende diejenige ausgibt, die den kleinsten MAE zur Folge hat. Diese war: Time Delay: 18; C: 400; epsilon: 0.3. Damit liegt der MAE bei 0.0189 (vorher 0.347). Die dazugehörige Grafik haben wir unter *3.1_stockPredict_optimized.png* abgespeichert. Die Grafik mit den vorgegebenen Werten haben wir noch unter *3.1_stockPredict.png* beigefügt. Zur besseren Übersichtlichkeit haben wir nur den letzten Abschnitt des Kurses ausgegeben.

[3] b102_stockMarketPrediction_evaluate.py