Applied Analytical Data Science

Teil 9: Zeitliche und räumliche Daten

Dr. Jörg-Uwe Kietz,
Vorlesung an der Univ. Zürich,

Mittwoch, 14:00-15:45 Uhr Vorlesung,
16:00-17:30 Uhr Übung

http://www.kietz.ch/AADS/

Heutiges Programm

- Data Mining und zeit abhängige Daten
  - Uni- und multivariate Zeitreihen
  - Einfache Zeitreihen-Analyse:
    - Glätten mit Durchschnittsbildung (Moving Averages),
    - Beispiel-Generierung mit gleitendem Fenster (Sliding Window),
    - Saisonale Effekte
    - Diskretisierung
  - Zeit abhängige Assoziations-Regeln
- Data Mining und räumliche Daten (spatial data)
- Zeitliche und räumliche Daten in ILP

Zeitreihen

- Univariate Zeitreihen: Messungen einer abhängigen Variablen über einem Zeitraum und in der Regel in festen Zeitabständen.
- Multivariate Zeitreihen: Messungen mehrerer abhängigen Variablen über dem gleichen Zeitraum und in denselben Zeitabständen.

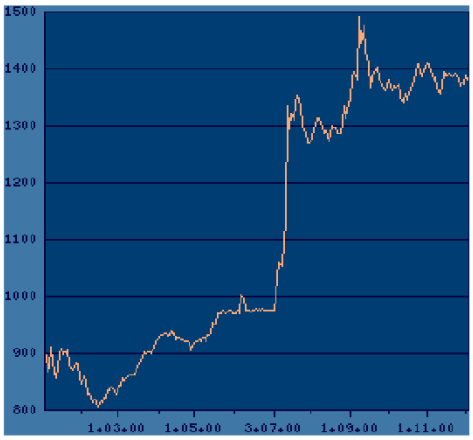
⇒ Unterschied zu bisherigen Datensätzen:
Die unabhängige Variable (die Zeit) ist monoton steigend, d.h. ein vergangener Zeitpunkt kehrt niemals wieder!

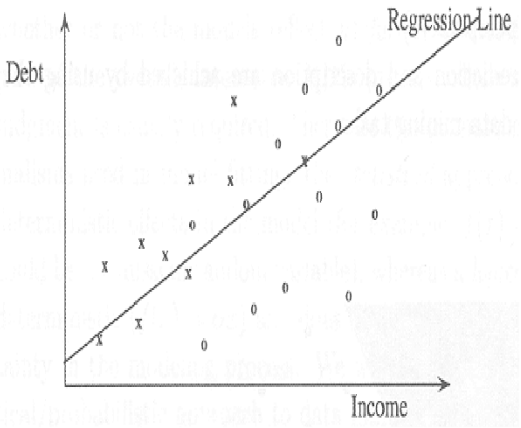
⇒ **Das Data Mining Ziel ist meist die Vorhersage des nächsten Wertes**

Univariate Zeitreihe	
Zeit i	Wert t(i)
1	v1
2	v2
3	v3
4	v4
5	v5
6	v6
7	v7
8	v8
9	v9
10	v10

Univariate Zeitreihe: Aktienkurse

Zeitreihen vs. Regression

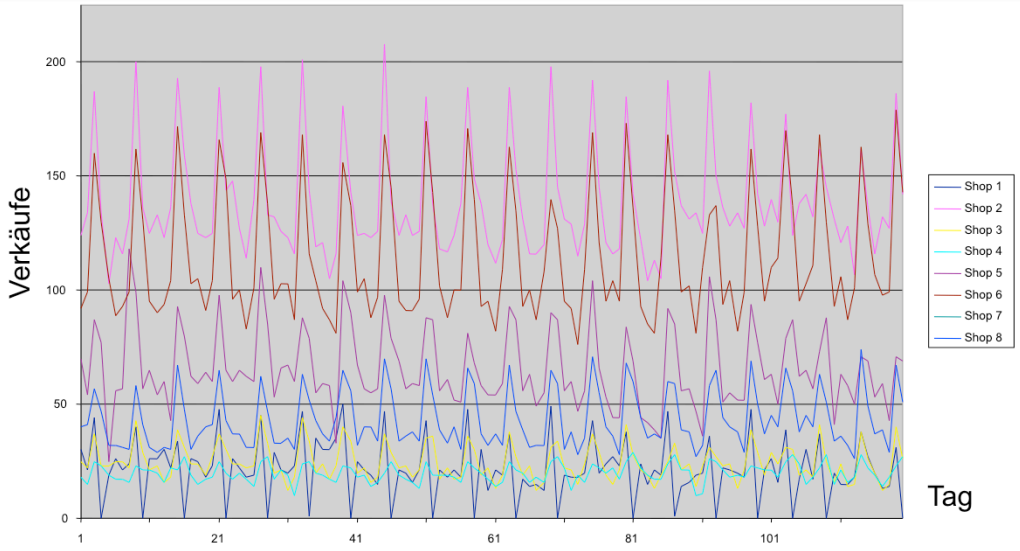




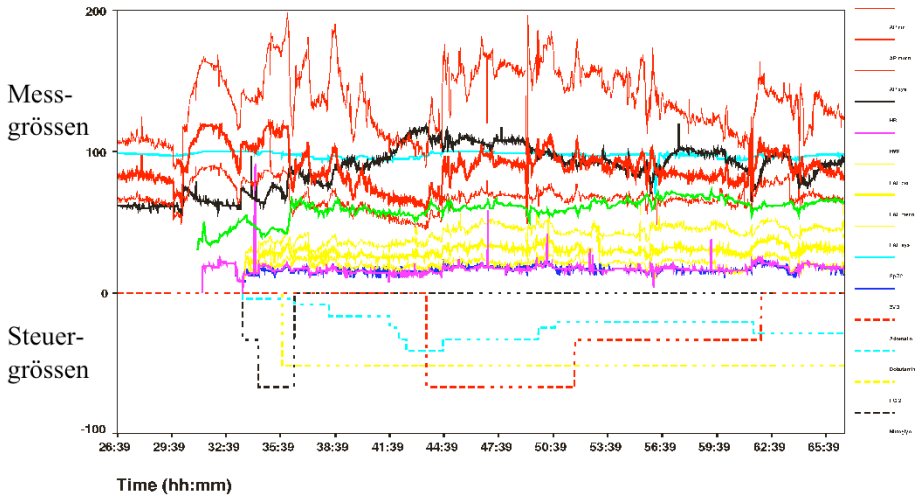
- Bei Zeitreihen kennen wir nur den linken Kontext
- Bei der Regression in der Regel beide Kontexte

Multivariate Zeitreihe: Aktienkurse

Multivariate Zeitreihe: Zeitungsverkäufe



Multivariate Zeitreihe: Prozesse & Steuerung



Heutiges Programm

- Data Mining und zeit abhängige Daten
  - Uni- und multivariate Zeitreihen
  - Einfache Zeitreihen-Analyse:
    - Glätten mit Durchschnittsbildung (Moving Averages),
    - Beispiel-Generierung mit gleitendem Fenster (Sliding Window),
    - Saisonale Effekte
    - Diskretisierung
  - Zeit abhängige Assoziations-Regeln
- Data Mining und räumliche Daten (spatial data)
- Zeitliche und räumliche Daten in ILP

Glätten durch Durschnittsbildung

- Die Idee des Glättens ist, zufällige Variationen und Störungen (Noise) durch Glätten der Übergänge von Nachbarpunkten auszugleichen
- Einfache Glättung (Moving Average):

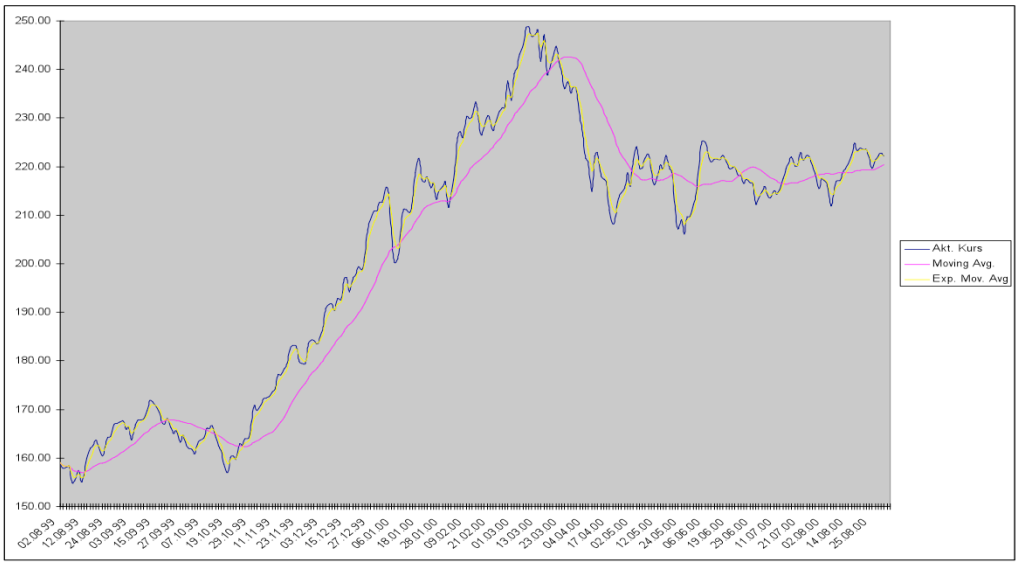
  $ma(i,m) = (1 / m) * \Sigma^{i}_{j=i - (m-1)} t(j)$

- Exponentielle Glättung (Exponential Moving Average):

  $ema(i,m) = p * t(i) + (1-p) * ema(i-1,m-1)$, für $m > 1$

  $t(i)$, für $m = 1$

Vergleich: Mov. Avg. und Exp. Mov. Avg.

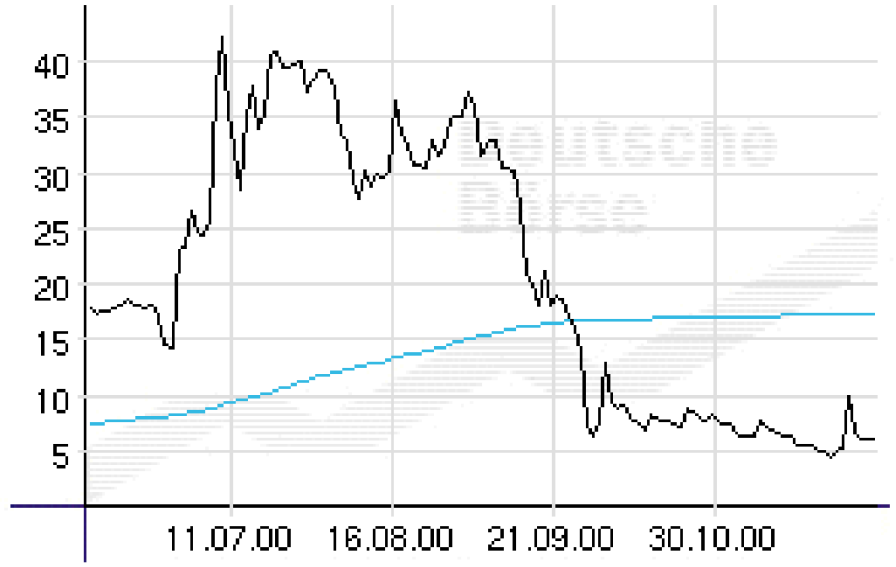


Beobachtungen zum Glätten

- Je mehr Zeitpunkte beim einfachen Glätten gemittelt werden, desto:
  - Glatter ist die Kurve
  - „Veralteter“ ist die Kurve
- Je kleiner **p** (die Gewichtung des aktuellen Zeitpunktes) beim Exponentiallen Glätten ist, desto:
  - Glatter ist die Kurve
  - „Veralteter“ ist die Kurve

⇒ Beim Glätten verliert man an Aktualität, d.h. man kann Trendwenden nur verfolgen, aber nicht vorhersagen.

Aktienkurs und 200 Tage Moving Average



Gleitendes Fenster (Sliding Window)

- Durch die Methode des gleitenden Fensters können Zeitreihen in Beispielmengen zum normalen Lernen umgewandelt werden:

Zeitreihe:

Zeit i	Wert t(i)
1	v1
2	v2
3	v3
4	v4
5	v5
6	v6
7	v7
8	v8
9	v9
10	v10

$\Rightarrow$

Beispielmenge (Grösse 5):

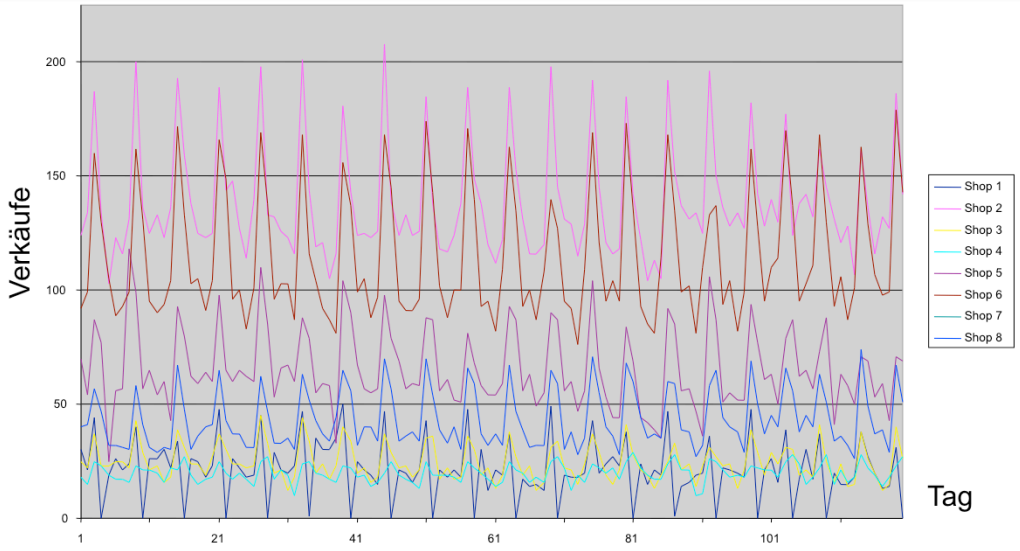
A1	A2	A3	A4	A5	Ziel
v1	v2	v3	v4	v5	v6
v2	v3	v4	v5	v6	v7
v3	v4	v5	v6	v7	v8
v4	v5	v6	v7	v8	v9
v5	v6	v7	v8	v9	v10

Die Fenstergrösse (Lag) bestimmt die Anzahl der erzeugten Attribute.

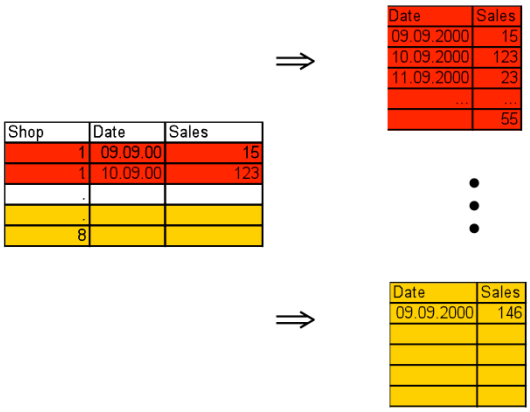
Gleitendes Fenster (Sliding Window)

- Über den erzeugten Beispielen kann mit ganz normalen Data Mining Methoden gelernt werden:
  - Regression, wenn das Zielattribut numerisch ist.
  - Klassifikation, wenn das Zielattribut nominal ist.

Ein Beispiel: Zeitungsverkäufe

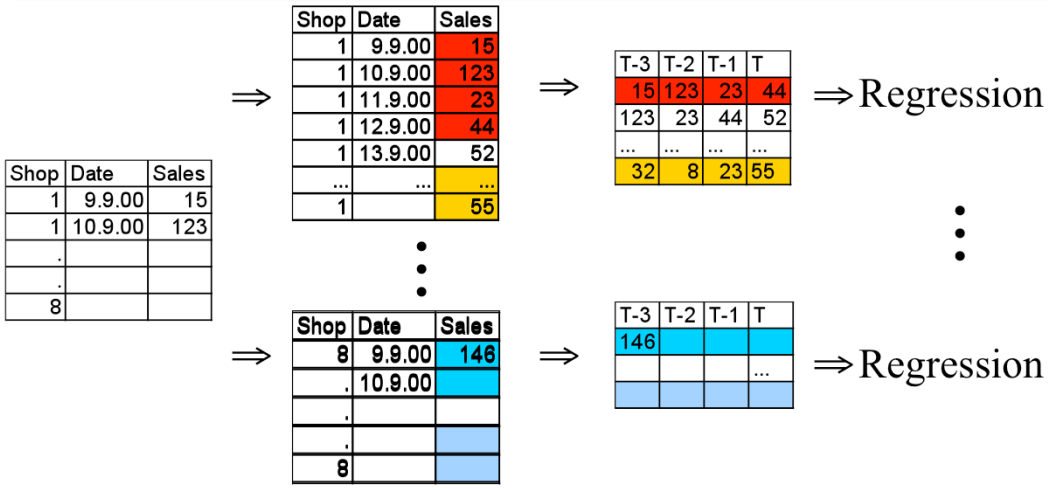


Vorhersage der Verkäufe des nächsten Tages



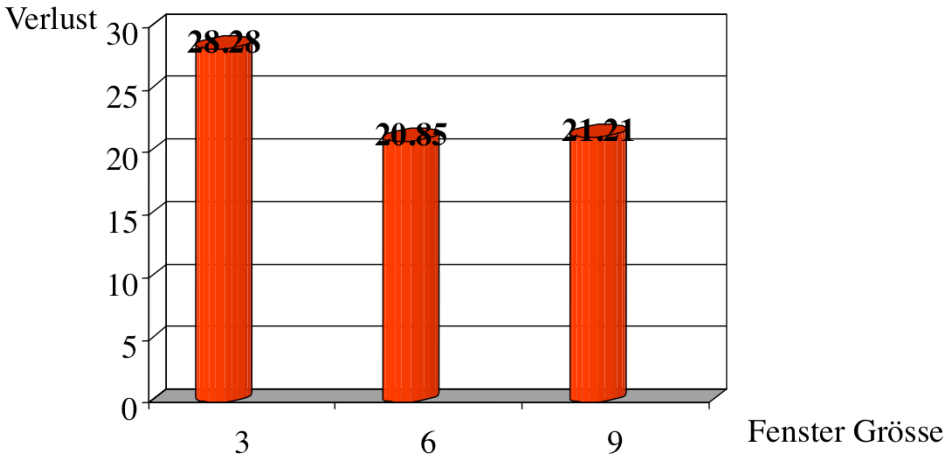
Extraktion der einzelnen Läden

Vorhersage der Verkäufe des nächsten Tages



Sliding Window (Grösse 3)

Einfluss der Fenstergrösse



⇒ Die Fenstergrösse hat starken Einfluss.

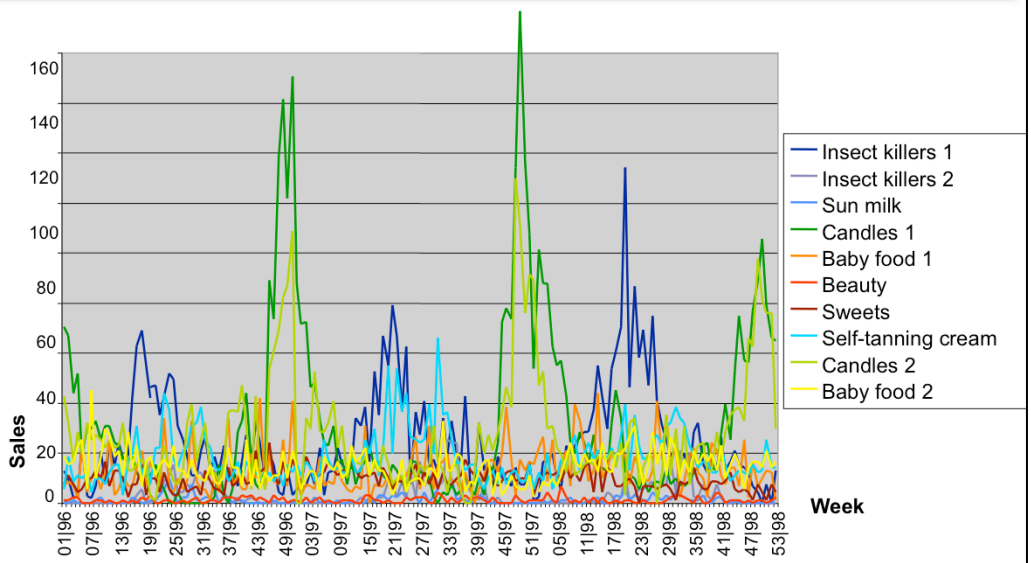
⇒ Sie kann oft nur experimentell optimiert werden.

Saisonale Effekte

- Saisonale Effekte sind
  - periodisch wiederkehrende (z.B. jedes Jahr)
  - Abschnitte (Saison, z.B. Weihnachtsverkäufe)

  bei denen sich das Verhalten der Zeitreihe deutlich von dem anderer Abschnitte (zu einer anderen Saison gehörig) unterscheidet.

Saisonale Effekte: Drogerie-Verkäufe

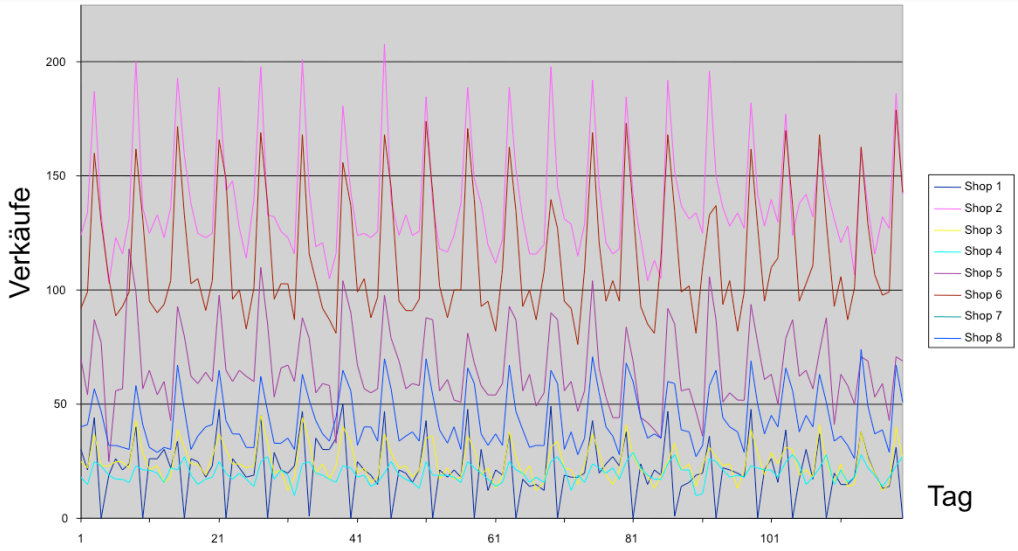


Behandlung saisonaler Effekte

- Man normalisiert die Zeitreihe und das Ergebnis durch einen saisonalen Faktor, z.B. den Mittelwert der letzten Weihnachtssaison.
- Man behandelt jede Saison als eigene Zeitreihe, z.B: die Folge aller Weihnachtsverkäufe.
- Man erweitert das Fenster, so dass zusätzlich auch Attribute zur letzten Saison vorhanden sind und diese von der Data Mining Methode berücksichtigt werden können.

$\Rightarrow$ Saisonale Effekte, ihre Periode und Dauer müssen erkannt und bei der Datenaufbereitung berücksichtigt werden.

Zeitungsverkäufe und Saison (Wochentag)



Vorhersage der Verkäufe pro Wochentag

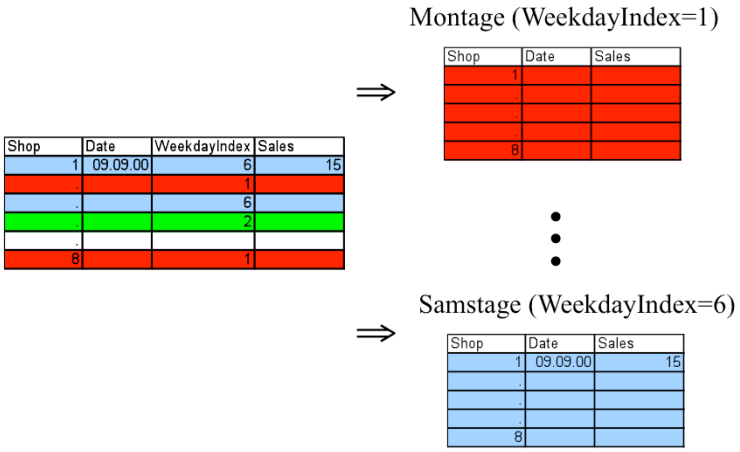
Shop	Date	Sales
1	09.09.00	15
.		
.		
.		
8		

⇒

Shop	Date	WeekdayIndex	Sales
1	09.09.00	6	15
.			
.			
.			
8			

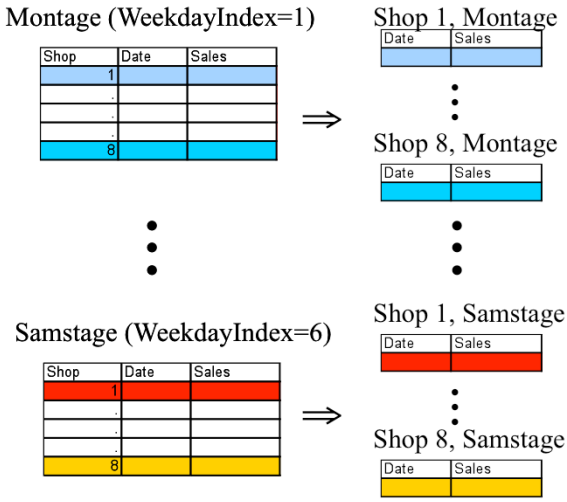
Konstruktion der Wochentag Attributs

Vorhersage der Verkäufe pro Wochentag



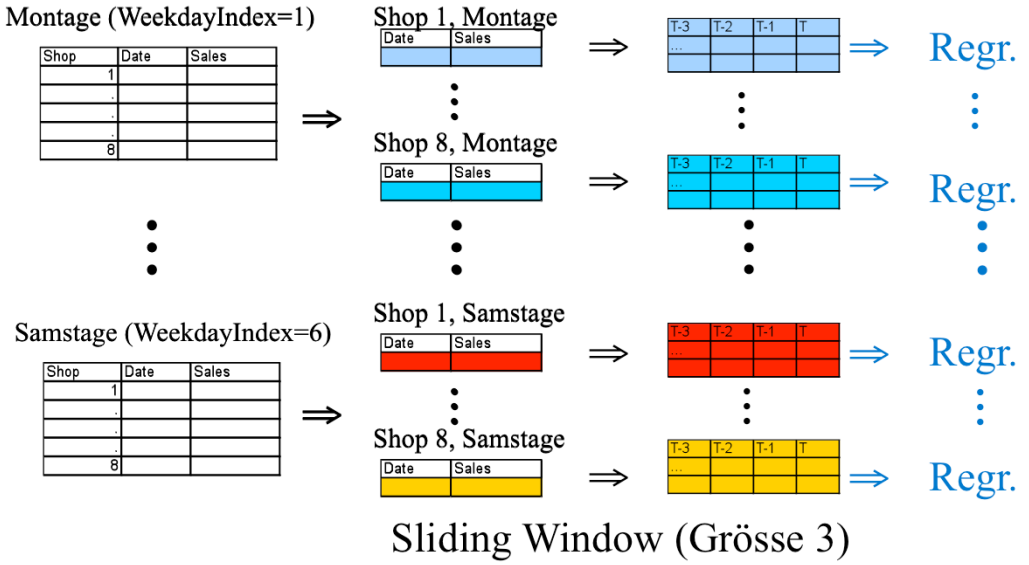
Segmentierung nach Wochentagen

Vorhersage der Verkäufe pro Wochentag



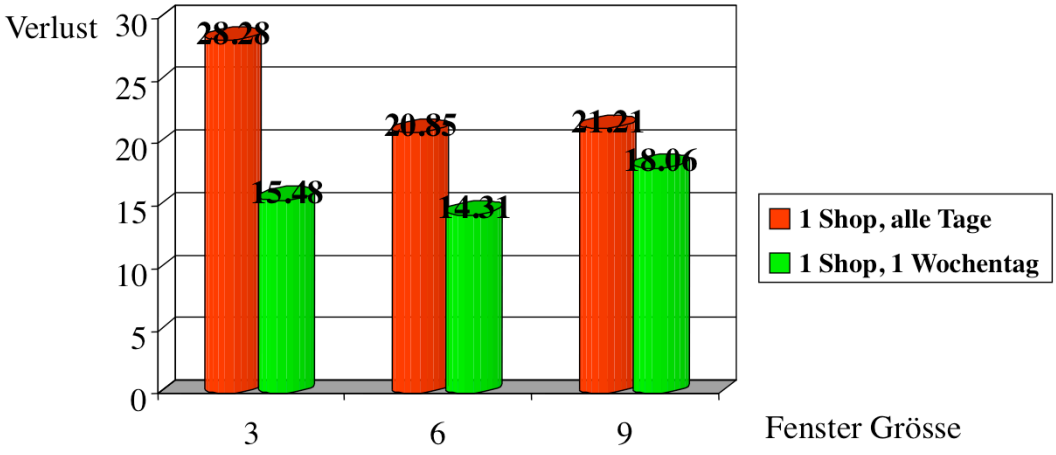
Extraktion der einzelnen Shops

Vorhersage der Verkäufe pro Wochentag



Sliding Window (Grösse 3)

Ergebnisse je nach Datenaufbereitung



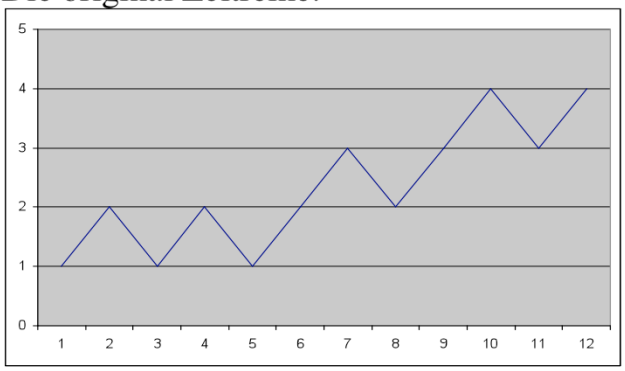
⇒ Fehler durch Saisonbehandlung um 14,9% - 45,3 % geringer!

Diskretisierung

- Zeitreihen können diskretisiert werden:
  - Data Mining Verfahren, die mit nominalen Daten (=> Klassifikation, Assoziationen, Subgruppenentdeckung) arbeiten, können angewendet werden.
  - Die diskretisierten Daten enthalten weniger zufällige Variationen und Störungen (Noise) (=> Glätten).
- Einflussgrössen der Diskretisierung:
  - Muster habe feste oder variable Grösse (=> Window).
  - Vorgegebene oder endeckte Muster (=> Clustering, Statistische Methoden) zur Diskretisierung.

Beispiel: Diskretisierung

Die original Zeitreihe:



Die Muster:

a1 = 

a2 = 

a3 = 

Die diskrete Zeitreihe (Window Grösse 3):

a1, a2, a1, a2, a3, a1, a2, a3, a1, a2

Heutiges Programm

- Data Mining und zeit abhängige Daten
  - uni- und multivariate Zeitreihen
  - Einfache Zeitreihenanalyse:
    - Glätten mit Durchschnittsbildung (Moving Averages),
    - Beispiel-Generierung mit gleitendem Fenster (Sliding Window),
    - Saisonale Effekte
    - Diskretisierung
  - Zeit abhängige Assoziations-Regeln
- Data Mining und räumliche Daten (spatial data)
- Zeitliche und räumliche Daten in ILP

Zeit abhängige Assoziationsregeln

- Definition: Eine zeit abhängige Assoziationsregel $A \Rightarrow^T B$ ist eine Assoziation über einer diskreten Zeitreihe, mit der Bedeutung: „Wenn A vorkommt, kommt B innerhalb der Zeit T vor.
- Der **Supportset(A,T)** einer Sequenz A ist die Menge aller Zeitfenster der Grösse T, in denen die Elemente der Sequenz $A = A_1,...,A_n$ in dieser Reihenfolge, aber nicht notwendigerweise aufeinanderfolgend, vorkommen.
- **Support($A \Rightarrow^T B$) = | Supportset((A,B),T) |**
- Die **Konfidenz($A \Rightarrow^T B$) = | Supportset((A,B),T) | / | Supportset(A,T) |**

Mining von zeit abhängigen Assoziationen

- Das Apriori-Verfahren (=> Teil 4) kann fast unverändert zum Finden von zeit abhängigen Assoziationsregeln benutzt werden, da Item-Mengen als sortierte Item-Listen implementiert werden.

=> Die Berücksichtigung der Sequenz (vs. Menge, insb. das Weglassen der Sortierung) macht Apriori zu Apriori-Time.

Aber: Apriori ist exponentiell in der Anzahl der verschiedenen Items, d.h. der Aufwand von Apriori ist $O(2^{|Items|})$

- Apriori-Time ist exponentiell in der Anzahl der möglichen Item-Sequenzen der Länge T, davon gibt es: $|Items|^T$

=> Der Aufwand von Apriori-Time ist $O(2^{|Items|^T})$

Heutiges Programm

- Data Mining und zeit abhängige Daten
  - uni- und multivariate Zeitreihen
  - Einfache Zeitreihen-Analyse:
    - Glätten mit Durchschnittsbildung (Moving Averages),
    - Beispiel-Generierung mit gleitendem Fenster (Sliding Window),
    - Saisonale Effekte
    - Diskretisierung
  - Zeit abhängige Assoziations-Regeln
- Data Mining und räumliche Daten (spatial data)
- Zeitliche und räumliche Daten in ILP

Räumliche Daten

- Datenbanken enthalten viele räumliche Daten:
  - Addressen, PLZ, Kanton, Land, ...
  - Zellen, Gemeinden, Regionen, ...
  - ....
- Diese Daten sind nominal und haben sehr viele verschiedenene Werte.

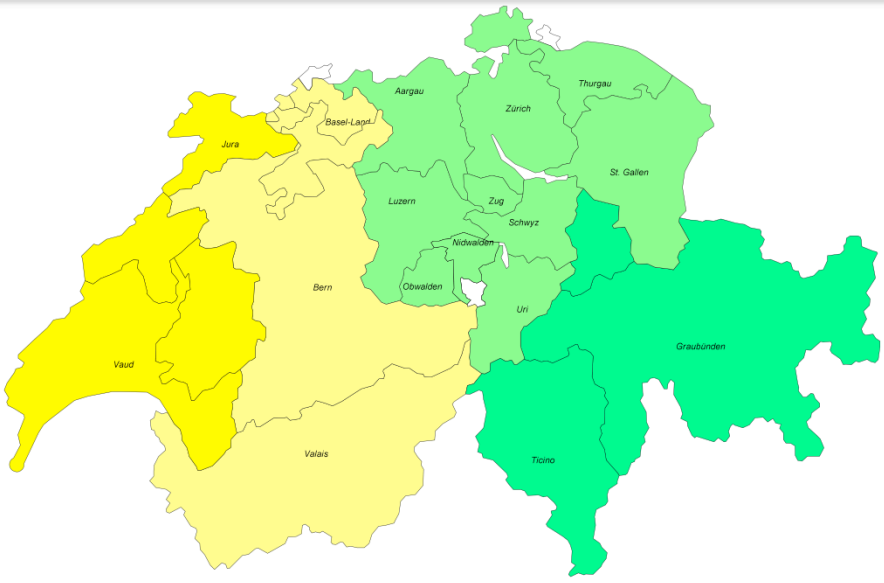
=> Sie sind zum Data Mining in dieser Form nicht geeignet.

=> Sie müssen in eine geeignete Form gebracht werden.

Nützliche räumliche Daten

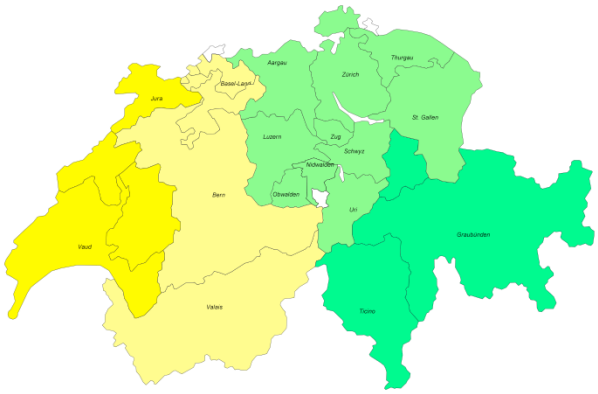
- Positionen in einem Koordinatensystem, z.B. Längen- und Breitengrade, d.h. skalare Attribute mit definierter Ähnlichkeit.
- Sehr abstrakte Gruppierungen, z.B. erste Ziffer der PLZ oder Kanton, im Bereich Schweiz, Länder im Bereich eines Kontinents, Kontinent im Bereich der Welt, ...
- Hintergrundwissen, d.h. viele besser geeignete Attribute die diese Orte beschreiben, welches mit Hilfe der Ortsangabe zugeordnet wird.
- Relationale Repräsentationen der Verbindungsnetze zwischen den Orten (Spatial Databases), z.B.: verbunden, benachbart, abstand, fahrt- oder flugzeit, ...

Visualisierung von Data Mining Ergebnissen



Erkennung von räumlichen Trends

- Menschen „sehen“ räumliche Trends.
- Die Erkennung von räumlichen Trends ist eine wichtige Data Mining Aufgabe:
  - Umweltdaten
  - Soziologische Daten
  - Wirtschaftsdaten
  - ....
- Erste experimentelle Data Mining Methoden



Heutiges Programm

- Data Mining und zeit abhängige Daten
  - uni- und multivariate Zeitreihen
  - Einfache Zeitreihen-Analyse:
    - Glätten mit Durchschnittsbildung (Moving Averages),
    - Beispiel-Generierung mit gleitendem Fenster (Sliding Window),
    - Saisonale Effekte
    - Diskretisierung
  - Zeit abhängige Assoziations-Regeln
- Data Mining und räumliche Daten (spatial data)
- Zeitliche und räumliche Daten in ILP

Zeitliche Daten in ILP

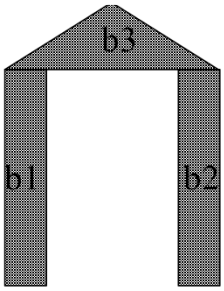
Zeit-Daten können einfach in Relationen dargestellt werden, z.B.

- visit(14,Mustermann, 110/80, ...), **previous_visit**(45,14)
  visit(45, Mustermann, 120/90, ...)
- measurement(trace3,**t1,t4**,left,60cm)
  measurement(trace3,**t4,t7**,left,80cm)
  increase_at(trace3,**t4**,left)

- Anwendungen:
  Therapievorschläge (MLT, GMD & FORTH), Roboter Sensor Daten (BLEARN, Univ Dortmund, Karlsruhe & Turin)

Räumliche Daten in ILP

Räumliche Daten können einfach in Relationen dargestellt werden, z.B.



on_top(b3,b1,b2),
touches(b3,b1),
distance(b1,b2,30m),
left_of(b1,b2), ...

- Anwendungen: Molekülstrukturen (King), MESH Design (Dolsak, 1988), VLSI-Design (Hermann, Beck 1994), Störungsdiagnose in Hochspannungsnetzen (MLT. 1993)

Fazit

- Zeitreihenanalyse beschäftigt sich mit der Vorhersage von zukünftigen Werten auf der Basis vergangener Werte.
- Windowing ist eine (Preprocessing-) Technik zur Anwendung von normalen DM-Tools (Klassifikation, Regression) auf Zeitreihen.
- Diskretisierung und Sequence-Assoziationsregeln sind eine weitere Möglichkeit Zeitreihen zu analysieren.
- Bei räumlichen (Spatial-) Daten ist das Konzept der räumlichen Nähe und Ausbreitung ein besonders betonter Einflussfaktor.
- Sowohl die Besonderheiten zeitlicher wie auch räumlicher Daten lassen sich mit multi-relationalen Verfahren (ILP) abdecken.

Literatur

Literatur zu Zeit:

- Weiss, S.; Indurkhya, N.: Predictive Data Mining, Morgan Kaufmann, 1998.
- Das, G.; Lin, K.; Mannila, H.; Renganathan, G.; Smyth, P.: Rule Discovery from time series, In: Proc of the Fourth Int. Conf. on Knowledge Discovery & Data Mining, KDD‘98, AAAI Press, 1998.

Literatur

Literatur zu Raum:

- Pyle, D.: Data Preparation for Data Mining, Morgan Kaufmann, 1999.
- Ester, M.; Frommelt, A.; Kriegel, H.; Sander J.: Algorithms for Characterisation and Trend Detection in Spatial Databases, In: Proc of the Fourth Int. Conf. on Knowledge Discovery & Data Mining, KDD'98, AAAI Press, 1998.