Eva SCHIRGI

Analyse von Messdaten des steirischen Luftgütemessnetzes durch funktionales Clustern

DIPLOMARBEIT

zur Erlangung des akademischen Grades einer
Diplom-Ingenieurin

Diplomstudium Technische Mathematik

Graz University of Technology

Technische Universität Graz

Betreuer:
Univ.-Prof. Dipl.-Ing. Dr.techn. Ernst STADLOBER

Institut für Statistik

Graz, im März 2012

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am ....
(Unterschrift)

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

....
date (signature)

Zusammenfassung

Luftgüte-Messungen stellen ein wichtiges Instrument für die Evaluierung von Maßnahmen zur Verbesserung der Luftqualität dar. In der Steiermark werden die Luftschadstoffe Feinstaub PM_{10} [µg/m³], Ozon O_3 [µg/m³], Schwefeldioxid SO_2 [µg/m³] und Stickstoffdioxid NO_2 [µg/m³] an insgesamt 52 Messstationen erhoben. Eine Einteilung der Messstationen aufgrund ähnlicher Eigenschaften soll Informationen zur Luftqualität in der Steiermark bringen, so dass möglicherweise eine räumliche Darstellung erreicht werden kann.

Der Ansatz für die Auswertung erfolgt über die Analyse der Tagesmittelwerte von Luftschadstoffkonzentrationen als funktionale (zeitabhängige) Daten. Die Beobachtungen von 24 Messstationen werden mithilfe von Basis-Spline-Funktionen (mit einem Knoten pro Monat) geglättet und mithilfe von funktionalen Clustern in Gruppen eingeteilt. Dabei werden zwei Methoden (*k-means*-Verfahren und *Partitioning Around Medoids*) angewandt, wobei die Anzahl der Cluster von 2 bis 4 variiert. Die Clusterergebnisse werden bzgl. verschiedener Charakteristiken (saisonale Abhängigkeiten, Konzentrationsniveau der Cluster) miteinander verglichen. Zusammenfassend kann man die Steiermark in einen nördlichen Bereich mit niedrigeren und in einen südlichen mit höheren Schadstoffkonzentrationen einteilen, wobei die höchsten Konzentrationen im Raum Graz auftreten.

Abstract

Air quality monitoring is an important tool in evaluation of air quality. In Styria, Austria there are 52 sites, which measure the air pollutants fine particulates (PM_{10} [µg/m³]), ozone (O_3 [µg/m³]), sulfur dioxide (SO_2 [µg/m³]) and nitrogen dioxide (NO_2 [µg/m³]). Classifying monitoring stations via homogeneous clusters should allow to gain more information about the air quality and the spatial patterns.

The daily means of air pollutant concentrations of 24 monitoring stations are considered as (time dependent) functional data, which will be smoothed by B-Spline-functions with one knot per month. Then the functional data are classified using functional cluster analysis, in particular by using the *k-means* algorithm and the *partitioning around medoids* (*PAM*) algorithm. Finally the clusterings are compared according to some characteristics (seasonal relationships, concentration levels of clusters). In summary it is possible to split Styria in two different parts: a northern region with rather low and a southern region with rather high air pollutant concentrations, where the highest concentrations appear in the region of Graz.

Danksagung

Ich möchte allen, die mich auf meinem Weg begleitet und unterstützt haben, insbesondere Herrn Univ.-Prof. Dipl.-Ing. Dr. techn. Ernst Stadlober, danken.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

I

EINLEITENDE WORTE

Mit der Richtlinie 2008/50/EG über Luftqualität und saubere Luft für Europa [EG 2008] sind für das europäische Gebiet grundlegende Bestimmungen, und damit Grenzwerte für die „Luftschadstoffe", festgelegt, die in Österreich als Immissionsschutzgesetz Luft (IG-L) [BGBL 1997] umgesetzt werden(vgl. FA 17C 2010/1, S. 4). Dieses Immissionsschutzgesetz Luft wurde 1997 erstellt, 2001 mit Grenzwerten für Feinstaub novelliert und 2007 aktualisiert (vgl. FA 17C 2010/2, S. 11). Die Ziele des Gesetzes sind der Schutz der Gesundheit des Menschen, des Tier- und des Pflanzenbestandes, der Schutz des Menschen vor unzumutbar belästigenden Luftschadstoffen, die Verringerung der Immission von Luftschadstoffen und die Bewahrung und Verbesserung der Luftqualität (vgl. FA 17C 2010/2, S. 11).

Zur Gewährleistung der Einhaltung der Grenzwerte bzw. der Umsetzung der festgelegten Maßnahmen zur Verringerung der Luftschadstoffe müssen mittlerweile regelmäßige Statuserhebungen erstellt und daraufhin die entsprechenden Maßnahmen gesetzt werden (vgl. FA 17C 2010/2, S. 9-11). In der Steiermark dient das Landes-Umwelt-Informationssystem (LUIS) dazu, „den Zustand der Umwelt (Gewässer, Luft, Boden, Tier- und Pflanzenwelt, natürliche Lebensräume, Lärm); Vorhaben und Tätigkeiten, die Gefahren für Menschen hervorrufen oder die Umwelt beeinträchtigen können" (http://www.umwelt.steiermark.at/), zu dokumentieren. Mit diesem System werden die Daten in den einzelnen Bezirken mit festen bzw. mobilen Messstationen erhoben.

Insbesondere wird auf die Luftschadstoffe Feinstaub (PM_{10}), Ozon (O_3), Schwefeldioxid (SO_2) und Stickstoffdioxid (NO_2) großes Augenmerk gelegt. Dazu werden andere Merkmale wie Temperatur, Windstärke bzw. -richtung, Sonneneinstrahlung, etc., erhoben. Die Daten werden halbstündlich von den Messstationen gemessen und gespeichert. Die Homepage des Magistrats Graz (http://www.umwelt.steiermark.at) bietet

die Möglichkeit, die Daten herunterzuladen, wobei die Daten entweder als Tages-, Monats- bzw. Jahresdurchschnittswerte aufbereitet und als Excel-Files zur Verfügung gestellt werden.

1 Daten

Die Daten stammen von 25 (der 52 steirischen Messstationen) und wurden über einen Untersuchungszeitraum von drei Jahren (von 2008 bis 2010) erhoben (siehe Abbildung 1.1.1). Nicht jede Messstelle liefert über den gesamten Untersuchungszeitraum und für alle vier Merkmale Messwerte, die Merkmale Feinstaub und Stickstoffdioxid werden 2010 von allen 25 Stationen erhoben, in den Jahren davor von 24. Die Ozonwerte werden bei zehn (von den ausgewählten 25) Messstationen erhoben, Schwefeldioxid von 19.

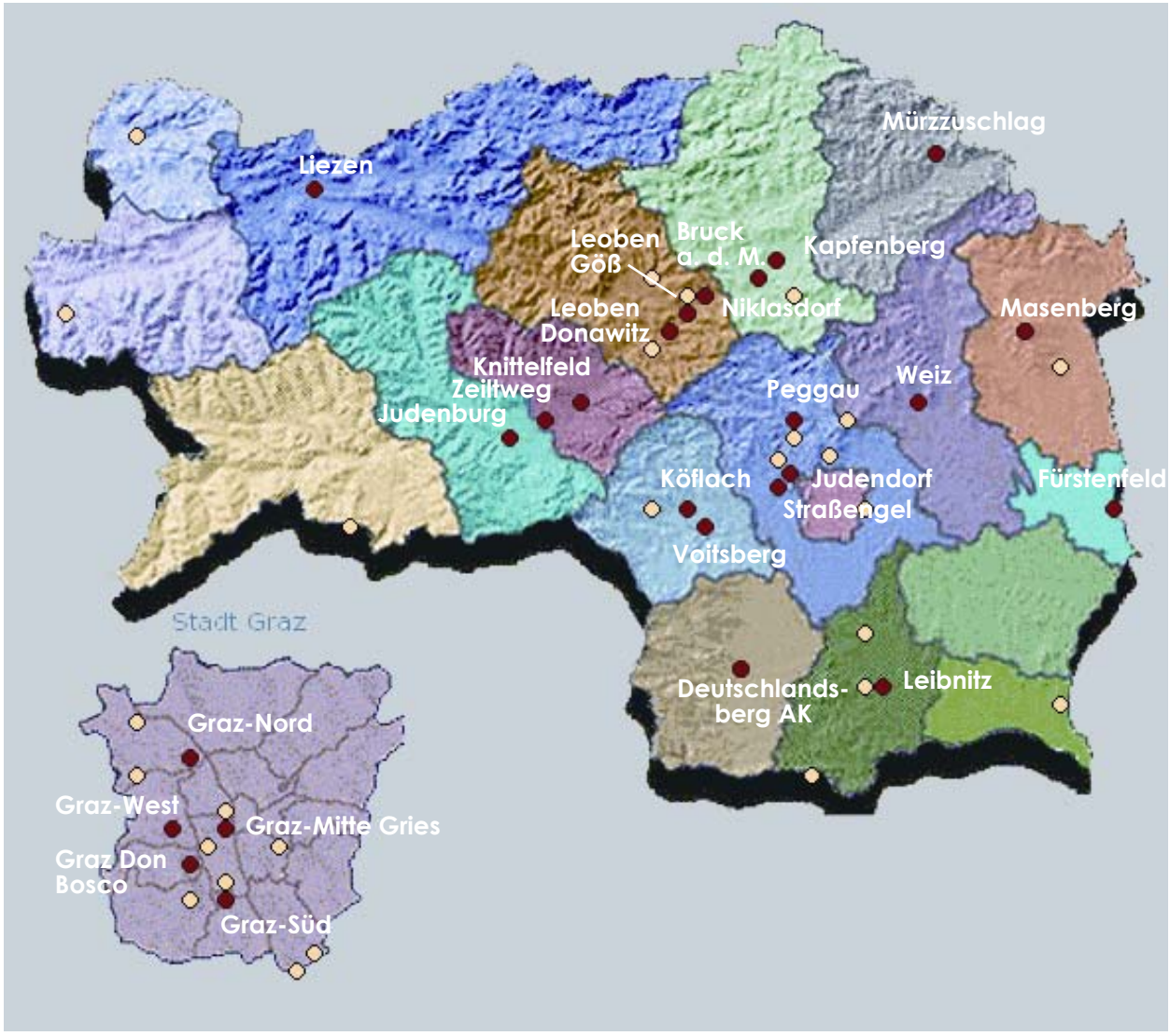


Abb. 1.1.1. Messstationen in der Steiermark [http://www.umwelt.steiermark.at/ am 4. 4. 2011, bearb.].

In drei steirischen Bezirken[1] (Feldbach, Murau und Radkersburg) gibt es keine Messstationen, die die erforderlichen Daten liefern, in acht Bezirken (Deutschlandsberg, Fürstenfeld, Hartberg, Knittelfeld, Leibnitz, Liezen, Mürzzuschlag und Weiz) jeweils eine. In Bruck, Judenburg und Voitsberg erfüllen jeweils zwei Messstellen die Kriterien für die Auswahl und in zwei Bezirken, nämlich in Graz-Umgebung und in Leoben, liefern drei Stationen die erforderlichen Daten. In der Stadt Graz gibt es fünf Messstellen, die die Luftschadstoffe messen, wobei Graz Mitte Gries erst ab Anfang 2010 Daten liefert.

2 Gliederung der Arbeit

Die vorliegende Arbeit ist in zwei Abschnitte geteilt, in einen theoretischen und einen empirischen Teil. Im ersten Teil werden die Begriffe und Methoden der funktionalen Datenanalyse und der Clusteranalyse, die für die Auswertung der Daten notwendig sind, eingeführt und erklärt.

Die Messwerte der steirischen Messstationen wurden über einen Zeitraum von drei Jahren erhoben. Man kann nun diese Daten multivariat betrachten, indem jede Beobachtung eine eigene Variable darstellt. Die funktionale Datenanalyse geht von der Annahme aus, dass den Beobachtungen ein funktionaler Zusammenhang zugrunde liegt. In Kapitel II wird eine Methode zur Schätzung einer (entsprechend glatten) Funktion aus vorhandenen Daten genauer beschrieben und zwar die Approximation mithilfe von Polynom-Splines, insbesondere durch Basis-Spline-Funktionen. Dabei wird das gesamte Intervall in Subintervalle zerlegt, auf denen Polynome vom Grad p geschätzt werden. An den Übergangsstellen sind noch Glattheitsbedingungen zu erfüllen. Diese Methode wird, gerade im Vergleich mit der Polynominterpolation, sehr geschätzt, da sie mit relativ geringem Rechenaufwand brauchbare Ergebnisse bzgl. ihrer Approximationseigenschaften liefert.

In Kapitel III werden die Clusteranalysemethoden veranschaulicht. Die Anwendung von Clusteranalyseverfahren ist insbesondere dann sinnvoll, wenn mehr als zwei Variablen simultan betrachtet werden müssen. Die Clusteranalyse besteht aus einer Reihe von Techniken, die bei einer gegebenen Datenmatrix eine Struktur aufdecken sollen. Erst die sinnvolle inhaltliche Interpretation der gefundenen Objekte-Cluster macht die

[1] Eine Tabelle mit der Aufteilung der Messstationen auf die steirischen Bezirke befindet sich in Anhang A.1.

Clusteranalyse erfolgreich. Somit ist es häufig sinnvoll, verschiedene Techniken hintereinander und unabhängig voneinander einzusetzen (Späth 1977, S. 12).

Mittlerweile gibt es viele Methoden, die aufgrund ihrer Voraussetzungen für unterschiedlich skalierte Daten verwendet werden. Eine Unterscheidung und Klassifizierung wird in Unterkapitel 1 vorgenommen. Danach werden Un- bzw. Ähnlichkeitsmaße, die die Grundlage für die Zuordnung von Daten zu den Gruppen darstellen, erklärt. Einige Verfahren, wie das *k-means*- bzw. *k-medoids*-Verfahren, werden genauer beschrieben, da diese Methoden für die Analyse der vorliegenden Daten herangezogen werden.

Im empirischen Teil werden die erhobenen Daten der Messstationen, und zwar die Luftschadstoffe Feinstaub (PM_{10}), Ozon (O_2), Schwefeldioxid (SO_2) und Stickstoffdioxid (NO_2) als Merkmale analysiert. Mithilfe der Splineexpansion werden die Zeitreihen als Funktionen in Abhängigkeit von der Zeit *t* dargestellt. Mithilfe der Clusteranalysemethoden sollen Gruppierungen von Messstationen gefunden werden, die Ähnlichkeiten im Verlauf der Schadstoff-Funktionen aufweisen. Für diese Analyse wird die Statistik Software *R* verwendet (http://www.cran.r-project.org, Version 2.14.1).

II

FUNKTIONALE DATENANALYSE

Ausgangspunkt für die funktionale Datenanalyse stellen Daten dar, die Veränderungen über eine stetige Variable, wie der Zeit, messen. Als Beispiel kann die Messung der Feinstaubpartikel in der Luft über einen Zeitraum von drei Jahren (von 1.1.2008 bis 31.12.2010) dienen. Diese Messwerte können z. B. mithilfe eines Liniendiagramms dargestellt werden (siehe Abbildung 2.1).

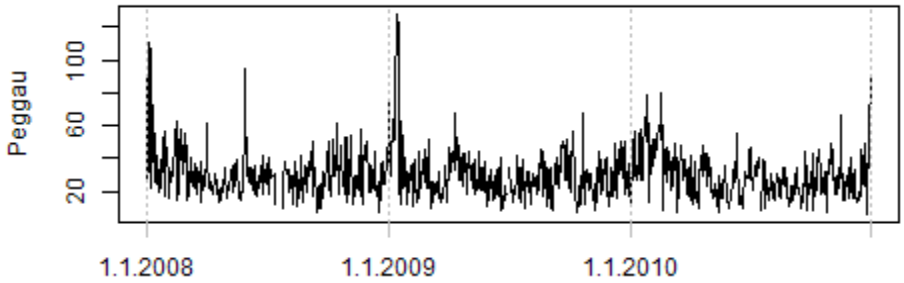


Abb. 2.1. Zeitreihenplot für das Merkmal PM_{10}

Man kann bereits einige erste Informationen ablesen, wie in diesem Fall, dass die Beobachtungen nicht zufällig verteilt sind, sondern jeweils von den vorherigen abhängig sind. In einem weiteren Schritt kann man zu der Annahme gelangen, dass den Daten eine stetige (und glatte) Funktion zugrunde liegt. Damit stellen die Messwerte „Momentaufnahmen" in Abhängigkeit von einer Kovariablen dar, wobei deren Veränderungen jedoch nicht sprunghaft, sondern kontinuierlich angenommen werden (vgl. Ramsay/Silverman 1997, S. 37). Die einzelne Beobachtung kann demnach als

$$y_j = x(t_j) + \varepsilon_j \qquad \varepsilon_j \overset{iid}{\sim} F(0,\sigma^2), \quad j = 1, \ldots, m$$

dargestellt werden, wobei t_j der Zeitpunkt der Messung und ε_j der zufällige unabhängige Fehler mit Verteilung F ist. Ziele der funktionalen Datenanalyse können unter ande-

rem die Aufbereitung der Daten für weitere Analysen, die Darstellung der Daten, um charakteristische Merkmale zu finden, das Erklären der Variationen in den Ergebnissen bzw. in den abhängigen Variablen, ein Vergleich von Datenmengen etc. sein (vgl. Ramsay/Silverman 1997, S. 8).

1 Definition und Anpassung funktionaler Daten

Ramsay und Silverman (1997) gehen von der grundlegenden Idee aus, die vorhandenen Daten als eine einzelne Struktur wahrzunehmen, und nicht als eine Folge von diskreten Beobachtungen. Eine funktionale Beobachtung besteht aus m Paaren (t_j, y_j), wobei y_j der Messwert der Funktion $x(t_j)$ ist (vgl. Ramsay/Silverman 1997, S. 37). Im Allgemeinen kann man davon ausgehen, dass man nicht nur eine einzelne funktionale Beobachtung x hat, sondern eine Sammlung von Funktionen x_i, $i = 1, \ldots, n$. Diese funktionalen Daten werden als m_i Paare (t_{ij}, y_{ij}), $j = 1, \ldots, m_i$, erhoben, wobei die Argumente t_{ij} für alle Funktionen x_i dieselben oder auch unterschiedlich sein können (vgl. Ramsay/Silverman 1997, S. 28). Zusätzlich wird vorausgesetzt, dass die Argumente t_{ij} über ein Intervall T festgelegt sind, und die Funktionen x_i bestimmten Stetigkeits- bzw. Glattheitsanforderungen genügen. Messungen sind meistens fehlerbehaftet, das heißt, dass eine reine Interpolation der (diskreten) Beobachtungen y_{ij} für eine Schätzung der Funktion x_i nicht ausreichend ist. Daher geht man vom Modell

$$y_{ij} = x_i(t_{ij}) + \varepsilon_{ij} \qquad \varepsilon_{ij} \overset{iid}{\sim} F(0,\sigma^2), \quad i = 1, \ldots, n, j = 1, \ldots, m_i$$

aus, wobei der Fehler oder die Störung ε_{ij} für die „Rauheit" der rohen Daten verantwortlich ist (vgl. Ramsay/Silverman 1997, S. 38).

Ferraty und Vieu (2006) gehen von einem allgemeineren Ansatz aus. Sie definieren eine funktionale (Zufalls-) Variable derart, dass sie Werte eines unendlich-dimensionalen Raumes annehmen kann. Beobachtungen dieser Variable werden dementsprechend als funktionale Daten bezeichnet. Eine funktionale Datenmenge besteht aus den Beobachtungen von mehreren funktionalen, identisch verteilten Variablen (vgl. Ferraty/ Vieu 2006, S. 6). Diese Definition kann auf viele Situationen angewandt werden, am häufigsten treten Funktionsmengen auf.

2 Glättung mithilfe der Basisfunktionen-Methode

Bei Glättungsverfahren für funktionale Daten ist nicht die Anzahl der Argumente t_j entscheidend, sondern die Krümmung der Funktion. Wenn die Krümmung stark ist, benötigt man eine größere Anzahl an Beobachtungen zur effektiven Schätzung der Funktion. Diese „Größe" wird in Abhängigkeit vom Zufallsfehler ε_j gesehen: Ist das Niveau des Fehlers gering und die Krümmung nicht sehr ausgeprägt, reicht eine relativ kleine Anzahl an funktionalen Daten (vgl. Ramsay/Silverman 1997, S. 39).

Es gibt unterschiedliche Möglichkeiten, um diskrete Beobachtungen y_j durch eine geeignete glatte Funktion x darzustellen (vgl. Ramsay/Silverman 1997, S. 39). Bei einem linearen Glättungsverfahren wird die Funktion x(*t*) durch eine Linearkombination von diskreten Beobachtungen der Form

$$\hat{x}(t) = \sum_{j=1}^{m} S_j(t) y_j$$

mit den Gewichten $S_j(t)$ geschätzt (vgl. Ramsay/Silverman 1997, S. 43). Einige der bekanntesten Glättungsverfahren beinhalten die Repräsentation der Funktion durch eine Linearkombination von *K* bekannten Basisfunktionen ϕ_k

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t).$$

Der Grad der Funktion ist abhängig von der Zahl der Basisfunktionen (vgl. Ramsay/ Silverman 1997, S. 44). Die Wahl der Basis entscheidet über die Anpassung und Qualität der Schätzung und die ($m \times K$) – Matrix $\Phi = (\phi_k(t_j))$ beinhaltet die Werte der Basisfunktionen.

Die Splinekoeffizienten c_k können mit dem Kleinste-Quadrat-Kriterium geschätzt werden. Die Zielfunktion SSE (sum of squared errors)

$$SSE = \sum_{j=1}^{m} \left[y_j - \sum_{k=1}^{K} c_k \phi_k(t_j) \right]^2$$

wird durch

$$\hat{c} = \left(\Phi^T \Phi \right)^{-1} \Phi^T y$$

minimiert, wobei $\hat{c}$ den *K*-dimensionalen Vektor der Splinekoeffizienten $\hat{c}_k$ darstellt und der Vektor *y* die beobachteten Werte enthält (vgl. Ramsay/Silverman 1997, S. 44).

In der Praxis werden für die funktionale Datenanalyse oft Fourier Basen für periodische Funktionen, polynomiale Basen, Spline-Basen oder Wavelet-Basen verwendet (vgl. Ramsay/Silverman 1997, S. 47-51). In weiterer Folge werden die Polynom-Splines genauer erklärt, da sie für die zu untersuchende Datenmenge eine gute Wahl darstellen.

2.1 Polynom-Splines

Polynom-Splines kommen in Bereichen zur Anwendung, wenn eine polynomiale Modellierung zur Darstellung der Daten nicht mehr ausreichend ist (vgl. Fahrmeir/Kneib/Lang 2007, S. 293f.). Oft geschieht es, dass selbst Polynome von hohem Grad (lokale) Maxima und Minima nicht entsprechend gut erfassen können. Wenn nun der Grad des Polynoms erhöht wird, werden zwar die Extremstellen besser approximiert, jedoch wird dabei an anderen Stellen der Daten die Schätzung ungenau (vgl. Fahrmeir/Kneib/Lang 2007, S. 294). Eine bessere Approximation erhält man durch stückweise polynomiale Funktionen. Dabei wird der Wertebereich (der Kovariablen) in Subintervalle zerlegt und separate Polynome für jedes Unterintervall ermittelt (vgl. Fahrmeir/Kneib/Lang 2007, S. 295). Zusätzlich ist es notwendig, Bedingungen für die Glattheit der Gesamtfunktion zu erfüllen, und zwar bei den Übergängen der separat geschätzten Polynomstücke. Diese Bedingungen führen zur Klasse der Polynom-Splines, wie sie von Fahrmeir et al. definiert werden (vgl. Fahrmeir/Kneib/Lang 2007, S. 295):

Definition 1. Polynom-Spline

Eine Funktion f: $[a, b] \rightarrow \mathbb{R}$ heißt Polynom-Spline vom Grad $p \geq 0$ zu den Knoten $a = \tau_1 < \ldots < \tau_K = b$, falls sie die folgenden Bedingungen erfüllt:

1. $f(z)$ ist $(p-1)$-mal stetig differenzierbar. Für $p = 1$ entspricht dies der Forderung, dass $f(z)$ stetig ist, für $p = 0$ werden keine Glattheitsanforderungen an $f(z)$ gestellt.
2. $f(z)$ ist auf den durch die Knoten gebildeten Intervallen $[\tau_j, \tau_{j+1})$ ein Polynom vom Grad p.

Durch die Forderung, dass die Polynomstücke $(p-1)$-mal stetig differenzierbar sind, wird die Glattheit bei den Übergängen garantiert. Der Grad der Splines ist für die globale Glattheit der Funktion entscheidend; so ist ein Polynom-Spline vom Grad 0 eine Stufenfunktion mit Unstetigkeitsstellen an den Knoten, vom Grad 1 ein Polygonzug bzw. stückweise lineare Funktion und vom Grad 2 eine stückweise quadratische Funktion. In der Praxis haben sich kubische Splines (Grad 3) bewährt.

Wenn die Daten keinen besonderen Funktionsverlauf erahnen lassen, sind äquidistante Intervalle für die Polynom-Splines gut geeignet. Die Anzahl der Knoten hat jedoch starken Einfluss auf die Funktionsanpassung. So wird die Approximation bei einer größeren Anzahl an Knoten flexibler, aber die (globale) Glattheit kann darunter leiden (vgl. Fahrmeir/Kneib/Lang 2007, S. 295).

Die einfachste Darstellung der Polynom-Splines ist eine Linearkombination der Basisfunktionen

$$\phi_k(t) = (t - T_k)_+^p = \begin{cases} (t - T_k)^p & t \geq T_k \\ 0 & sonst \end{cases} \quad k = 1, \ldots, K$$

wobei hierbei nur die $(K - 2)$ inneren Knoten verwendet werden. Diese Basen werden abgeschnittene Potenzen genannt und in einigen speziellen Anwendungen verwendet (vgl. Ramsay/Silverman 1997, S. 49). In den meisten Fällen sind dagegen die sogenannten Basic-Splines-Basen besser geeignet.

2.2 Basic-Spline-Basisfunktionen

Eine Möglichkeit zur Darstellung von Polynom-Splines bieten die Basic-Spline-, kurz B-Spline-, Basisfunktionen an. Eine B-Spline-Basis besteht aus $(p + 1)$ Polynomstücken vom Grad p, die an den Knoten $(p - 1)$-mal stetig differenzierbar zusammengesetzt sind. Ist die Knotenmenge äquidistant verteilt, dann haben die mittleren Polynomstücke dieselbe Form, ansonsten unterscheiden sie sich (vgl. Fahrmeir/Kneib/Lang 2007, S. 303f.). Abbildung 2.2.1 zeigt vollständige B-Spline-Basen sowohl für eine gleichmäßige als auch für eine ungleichmäßige Verteilung der Knoten.

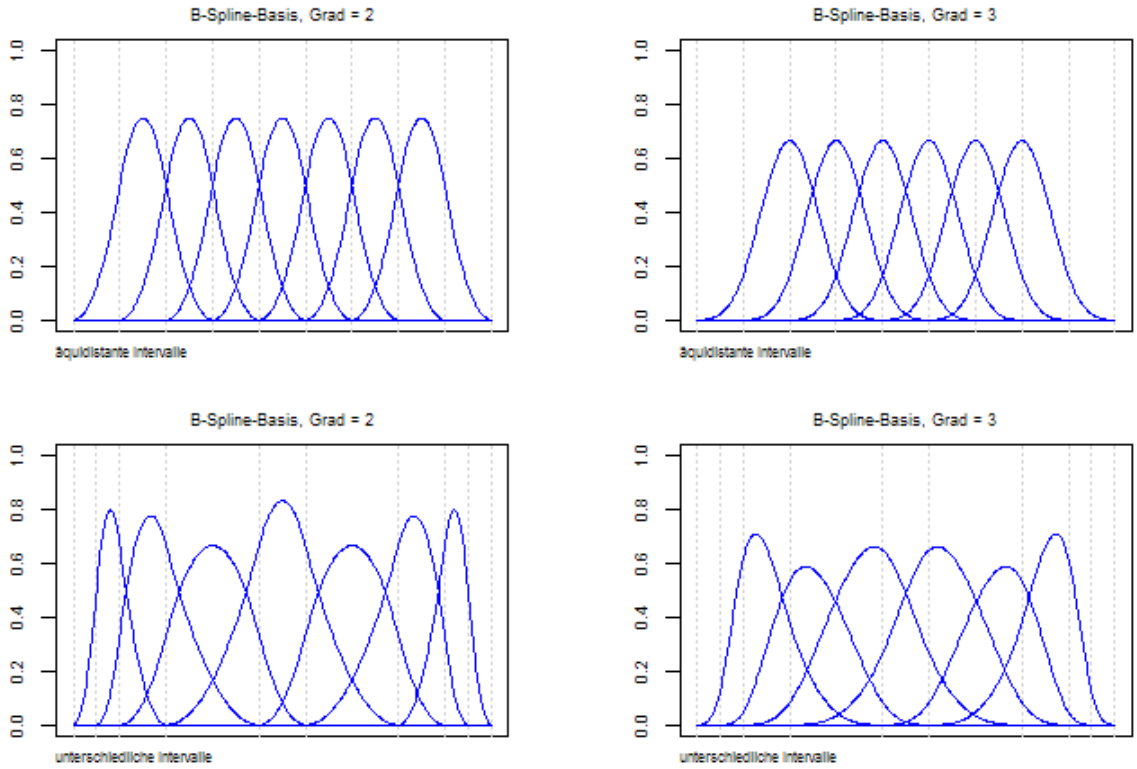


Abb. 2.2.1. Darstellung der B-Spline-Basen vom Grad 2 und 3, wobei die Knoten äquidistant (obere Reihe) und ungleichmäßig (untere Reihe) verteilt sind.

Die diskreten Messwerte y_{ij} können als

$$y_{ij} = x_i(t_j) + \varepsilon_{ij} \qquad \varepsilon_{ij} \overset{iid}{\sim} F(0,\sigma^2), \quad i = 1, \ldots, n, j = 1, \ldots, m_i$$

dargestellt werden, wobei x_i entsprechend glatte Funktionen über die Zeit t_j sind und ε_{ij} unabhängige Zufallsfehler mit Verteilung F. Mithilfe der vollständigen Basen kommt man zu einer Darstellung von $x_i(t)$ als Linearkombinationen der $B = K + p - 1$ Basisfunktionen, und zwar

$$x_i(t) = \sum_{k=1}^{B} \beta_{ik} B_k(t), \quad i = 1, \ldots, n$$

(Fahrmeir/Kneib/Lang 2007, S. 303f.). Dazu müssen die Spline-Funktionen auf der Knotenmenge $K = \{\tau_1, \ldots, \tau_m\}$ mit $a = \tau_1 < \ldots < \tau_m = b$ definiert werden

$$s(.,\beta_i) = \sum_{k=1}^{B} \beta_{ik} B_k(.), \quad i = 1, \ldots, n$$

mit β_i als Vektor der Splinekoeffizienten und $(B_1, \ldots, B_{K+p-1})$ der Basis der Spline-Funktionen (vgl. Ignaccolo/Ghigo/Giovenali 2008, S. 676). Aufgrund dieser Darstellung können die Splinekoeffizienten mit

$$\hat{\beta}_i = \left(\hat{\beta}_{i1}, \ldots, \hat{\beta}_{iB}\right) = \arg\min_{\beta_i} \frac{1}{m} \sum_{j=1}^{m} \left(x_{ij} - s(t_j; \beta_i)\right)^2$$

und in weiterer Folge die Funktionen $x_i(t)$ der Form

$$\hat{x}_i(t) = \sum_{k=1}^{B} \hat{\beta}_{ik} B_k(t), \quad i = 1, \ldots, n$$

geschätzt werden (vgl. Ingaccolo/Ghigo/Giovenali 2010, S. 676). Da der analytische Ausdruck der B_k's bekannt ist, können die Ableitungen exakt berechnet werden, d.h. man kann die geschätzten Funktionen leicht differenzieren

$$\hat{x}_i^{(r)}(t) = \sum_{k=1}^{B} \hat{\beta}_{ik} B_k^{(r)}(t) \quad i = 1, \ldots, n,$$

wobei der geschlossene Ausdruck von $B_k^{(r)}(t)$ bekannt ist (vgl. Ferraty/Vieu 2006, S. 32; vgl. Fahrmeir/Kneib/Lang 2007, S. 305).

Ein wesentlicher Vorteil der B-Spline-Basisfunktionen liegt darin, dass sie nur lokal definiert sind. Somit ist jede Basisfunktion nur in einem Bereich von $p + 2$ Knoten positiv. Bei einer Wahl von äquidistanten Knoten haben die Basisfunktionen die gleiche Form und an jeder beliebigen Stelle überlappen sich genau $p + 1$ Basisfunktionen. Zusätzlich ergibt sich, dass der Wertebereich der einzelnen Basisfunktionen nach oben beschränkt ist (vgl. Fahrmeir/Kneib/Lang 2007, S. 305).

Bei einer nichtparametrischen Funktionsschätzung entscheidet die Anzahl der Knoten über die Qualität der Polynom-Splines. Dafür gibt es im Großen und Ganzen zwei grundlegende Strategien, d.h. eine Regulierung kann einerseits durch eine adaptive Wahl der Knoten und andererseits durch Penalisierungsansätze erreicht werden (vgl. Fahrmeir/ Kneib/Lang 2007, S. 307).

Bei den penalisierten Splines, auch P-Splines genannt, wird die Funktion $x(t)$ im ersten Schritt durch ein Polynom-Spline geschätzt, basierend auf einer großen Anzahl an Knoten (üblicherweise zwischen 20 und 40 Knoten), wodurch die Funktion recht gut approximiert wird. Im nächsten Schritt wird ein Strafterm eingeführt, der eine zu große Variabilität der Schätzung verhindert (vgl. Fahrmeir/Kneib/Lang 2007, S. 308). Für die Approximierung einer Funktion mithilfe einer B-Spline-Basis wird häufig ein Strafterm verwendet, der auf der zweiten Ableitung basiert, die die Krümmung einer Funktion definiert und einen Glättungsparameter λ enthält,

$$\lambda \int (x''(t))^2 dt .$$

Dieser Ansatz wird auch von der Statistik-Software *R* (Paket *mgcv*) verwendet. Fahrmeir et al. (2007) geben einen einfacheren Penalisierungsansatz bei äquidistanten Knoten mithilfe von Differenzen an, da die erste Ableitung eines B-Splines mithilfe der ersten Differenzen der zugehörigen Parameter berechnet werden kann (vgl. Fahrmeir/Kneib/ Lang 2007, S. 309f.).

III

CLUSTERANALYSE

Mithilfe der Clusteranalyse soll eine Struktur in einer Menge von Objekten aufgedeckt bzw. erzeugt werden, wobei die Beobachtungsmerkmale für die Klassifizierung verantwortlich sind. Ziel ist es, dass sich die Objekte innerhalb einer Gruppe bzgl. der Merkmalsausprägungen sehr wenig unterscheiden, aber zu anderen Gruppen unähnlich sind (vgl. Pruscha 2006, S. 289). Die Zuordnung der Objekte zu den Clustern ist dabei nicht immer eindeutig, wie man in Abbildung 3.1 sehen kann. Einerseits entscheidet die Anzahl der Gruppen über das Aussehen der (gefundenen) Struktur, wobei die kleinstmögliche Clusterung aus einem Cluster aller Objekte und das größtmögliche aus *n* Cluster mit jeweils einem Element besteht. Andererseits bleibt nach der Festlegung der Anzahl der Gruppen noch immer die Frage, wie die Elemente den einzelnen Gruppen zugeordnet werden sollen (siehe Abbildung 3.1, Graphik 3 und 4).

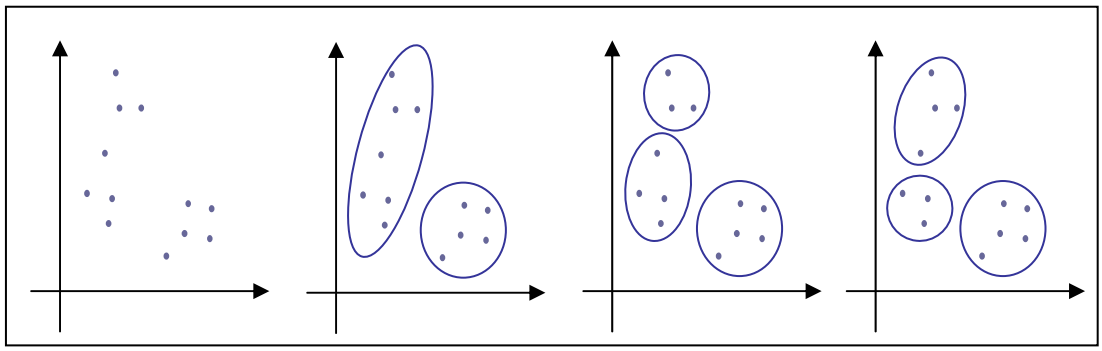


Abb. 3.1. Clusterung einer Datenmenge mit $k = 2$ bzw. $k = 3$

Im Gegensatz zu Methoden der induktiven Statistik wird die Clusteranalyse als deskriptives bzw. exploratives Instrument eingesetzt, das bedeutet, dass die gefundene Struktur die Grundlage für andere Analysemethoden, wie Diskriminanzanalyse oder Varianzanalyse darstellt (vgl. Pruscha 2006, S. 289; vgl. Kaufman/Rousseeuw 1990, S. 37).

Somit ist sehr wohl entscheidend, welche Clusterverfahren angewandt werden und welche Strukturen sie abbilden.

1 Begriffe und Methoden der Clusteranalyse

Der Ausgangspunkt für eine Clusteranalyse sind *n* Fälle (Objekte) mit *p* Merkmalen (Variablen), die in einer $n \times p$ – Datenmatrix zusammengefasst werden (siehe Tabelle 3.1.1). Diese Matrix bildet die Grundlage für die weitere Clusteranalyse (vgl. Pruscha 2006, S. 289).

Objekt	Merkmal 1	Merkmal 2		Merkmal p
1 (x_1)	x_{11}	x_{12}	…	x_{1p}
2 (x_2)	x_{21}	x_{22}	…	x_{2p}
⋮				
n (x_n)	x_{n1}	x_{n2}	…	x_{np}

Tab. 3.1.1. Datenmatrix bei der Clusteranalyse

Unter der Struktur einer Datenmenge versteht man bei der Clusteranalyse eine Aufteilung der Objekte {*1, 2, …, n*} in *k* Klassen. Mengentheoretisch kann das als eine Partition

$$\mathcal{A} = (A_1, A_2, \ldots, A_k)$$

der Menge der *n* Objekte dargestellt werden, wobei eine nichtleere Menge A_i mit $|A_i| \geq 1$ genau einem Cluster entspricht. Im Optimalfall gibt die Partition die Struktur der Daten genau wieder, das heißt, dass die Objekte eindeutig zugeordnet sind. Theoretisch könnte man durch Enumeration diese optimale Lösung finden, jedoch ist die Anzahl der möglichen Partitionen bereits bei einer kleinen Menge an Objekten und bei wenigen Klassen sehr groß (vgl. Pruscha 2006, S. 291).

1.1 Klassifizierung der Clusteranalyseverfahren

Es gibt mittlerweile sehr viele Clusteranalyseverfahren, die nach unterschiedlichen Gesichtspunkten systematisiert werden. Ein erstes Kriterium der Unterscheidung ist die algorithmische Betrachtung (siehe Abbildung 3.1.1), bei der zwischen hierarchischen und nicht-hierarchischen (partitionierenden) Verfahren unterschieden wird (vgl. Kaufman/Rousseeuw 1990, S. 38-50).

Bei der partitionierenden Methode wird die Anzahl der Cluster *k* vorgegeben und der Algorithmus bestimmt die Zuordnung der Klassifikationsobjekte. Die Schwierigkeit be-

steht darin, dass nicht jede Klassenanzahl eine „gute“ Clusterung ergibt. Für die Praxis bedeutet das, dass der Algorithmus für verschiedene *k* durchlaufen wird und die Entscheidung für eine Aufteilung mithilfe von Gütekriterien erfolgt (vgl. Kaufman/ Rousseeuw 1990, S. 38f.).

Bei den hierarchischen Verfahren unterscheidet man zwischen agglomerativen und divisiven Verfahren, abhängig davon, ob man von einem Cluster mit allen Klassifikationsobjekten (divisiv) oder von *n* Clustern mit jeweils einem Objekt (agglomerativ) ausgeht. Die Idee besteht darin, dass alle möglichen *k* Partitionen, *k = 1, …, n*, gebildet werden, wobei in einem Schritt entweder zwei „ähnliche“ Cluster miteinander verschmolzen werden (divisives Verfahren) oder ein Cluster in zwei aufgeteilt wird (agglomeratives Verfahren) (vgl. Kaufman/Rousseeuw 1990, S. 44).

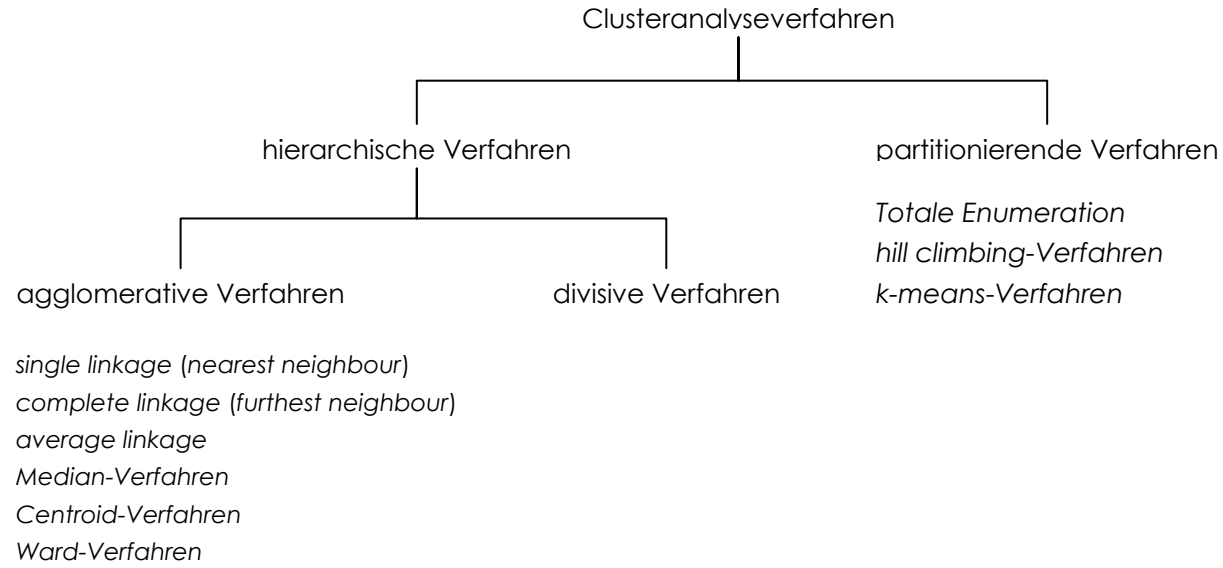


Abb. 3.1.1. Übersicht über Verfahren der Clusteranalyse nach algorithmischen Überlegungen

Bacher (1996) nimmt eine Klassifikation nach der Art der Zuweisung der Klassifikationsobjekte zu den Clustern vor (siehe Abbildung 3.1.2) und unterscheidet zwischen unvollständigen, deterministischen und probabilistischen Verfahren (vgl. Bacher 1996, S. 4f.). Bei der unvollständigen Clusterung steht die räumliche Darstellung der Klassifikationsobjekte im Vordergrund. Der Begriff der „Unvollständigkeit“ ergibt sich aufgrund der Tatsache, dass die Bildung der Cluster und die Zuordnung der Objekte bei der Interpretation durch den Anwender geschehen (vgl. Bacher 1996, S. 4f.).

Im Gegensatz dazu werden bei deterministischen Clusteranalyseverfahren die Klassifikationsobjekte den Clustern mit der Wahrscheinlichkeit 1 abhängig vom gewählten Ähnlichkeits- oder Unähnlichkeitsmaß zugeordnet (vgl. Bacher 1996, S. 4f.). Diese Einteilung kann eindeutig sein, d.h., dass sich jedes Objekt nur in einem einzelnen Cluster

befindet (disjunktive Clusteranalyse), oder auch mehrdeutig im Falle der überlappenden Clusteranalyse.

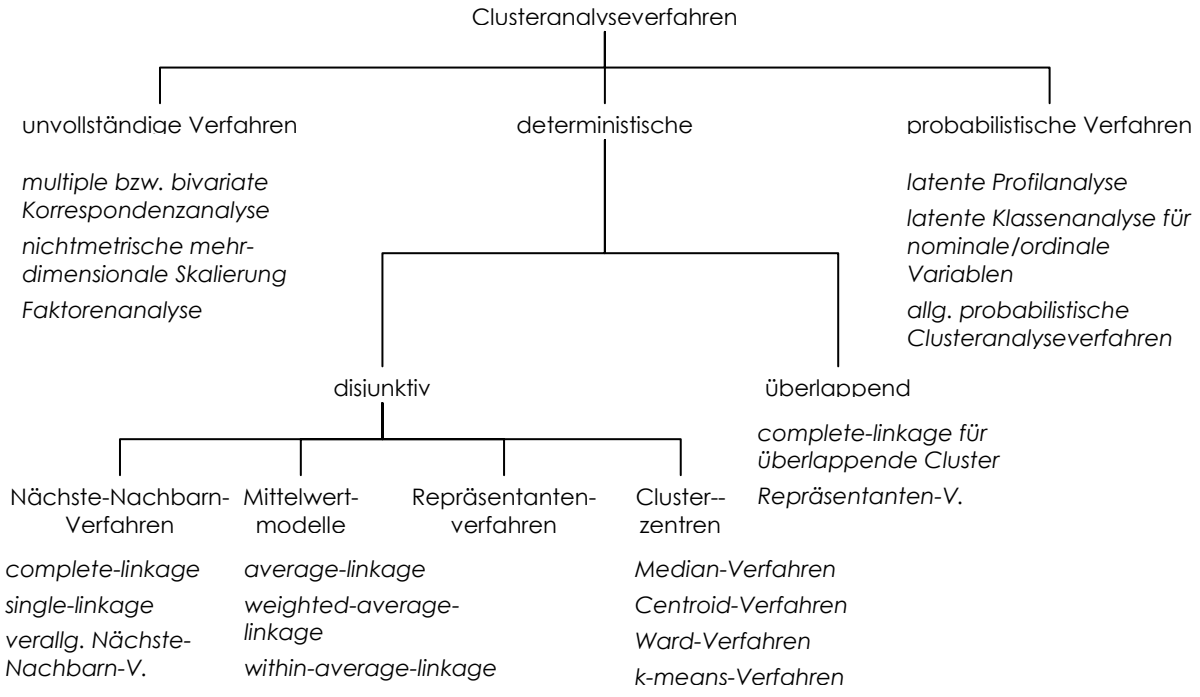


Abb. 3.1.2. Einteilung der Clusteranalyseverfahren nach Bacher (1996)

Disjunktive deterministische Verfahren können mithilfe der Vorstellung über die zu bildenden Cluster nochmals in Nächste-Nachbarn-Verfahren, Mittelwerteverfahren, Repräsentanten-Verfahren und hierarchische Verfahren zur Konstruktion von Clusterzentren unterteilt werden (vgl. Bacher 1996, S. 142f.). Beim Nächste-Nachbarn-Verfahren hat jedes Klassifikationsobjekt eine bestimmte Anzahl von nächsten Nachbarn (*complete linkage*, *single linkage*, *verallgemeinertes Nächste-Nachbarn*-Verfahren) oder einen *b*-ten nächsten Nachbarn (vgl. Bacher 1996, S. 142f.). Bei den Mittelwert-Modellen werden die Klassifikationsobjekte innerhalb eines Clusters durch durchschnittliche paarweise Ähnlichkeit bzw. Unähnlichkeit charakterisiert. Dazu zählen *average-*, *weighted average-* und *within-average-linkage* (vgl. Bacher 1996, S. 143).

Das grundlegende Element der Repräsentantenverfahren ist die Suche nach einem typischen Klassifikationsobjekt, das das Cluster repräsentieren soll. Die restlichen Objekte werden dann aufgrund ihrer Ähnlichkeit zu diesem Repräsentanten zugeordnet oder bleiben den Gruppierungen aufgrund der Unähnlichkeit unklassifiziert (vgl. Bacher 1996, S. 143). Eine weitere Möglichkeit der Clusterbildung kann erfolgen, indem nicht einzelne Objekte als Repräsentanten gewählt werden, sondern sogenannte Clusterzentren, die mithilfe der Mittelwerte der Variablen gebildet werden. Die Cluster können dann

einerseits über die maximale Entfernung der Clusterzentren oder andererseits über die maximale Streuung der Clusterzentren bestimmt werden (vgl. Bacher 1996, S. 143). Das *Median-*, *Zentroid-*, *Ward-* und die *k-means*-Verfahren basieren auf dieser Methode.

Bei den probabilistischen Methoden werden die Klassifikationsobjekte den Clustern mit einer Wahrscheinlichkeit zwischen 0 und 1 zugewiesen (vgl. Bacher 1996, S. 5). Bacher (1996) zählt dazu die latente Profilanalyse, die latente Klassenanalyse für nominale und ordinale Variablen und allgemeine probabilistische Clusteranalyseverfahren (vgl. Bacher 1996, S. 5f.).

1.2 Unähnlichkeits- und Ähnlichkeitsmaße

Die Zuordnung der Klassifizierungsobjekte zu einem (oder auch mehreren) Clustern erfolgt bei den meisten Clusteranalyseverfahren über das Kriterium der „Ähnlichkeit" bzw. der „Unähnlichkeit". Es gibt mehrere Möglichkeiten, diese Kriterien mithilfe von Funktionen zu quantifizieren. Bacher (1996) nennt in diesem Zusammenhang vier Gruppen: (i) Korrelationskoeffizienten als Maß für die Ähnlichkeit, (ii) Distanzmaße für die Darstellung der Unähnlichkeit, (iii) durch monotone Transformationen aus Distanzmaßen bzw. Korrelationskoeffizienten abgeleitete Un- und Ähnlichkeitsmaße und (iv) Un- und Ähnlichkeitsmaße, die für spezifische Fragestellungen eingeführt wurden. Die Wahl des entsprechenden Maßes hängt von der Zielsetzung der Clusteranalyse ab. Grundsätzlich sollen Korrelationskoeffizienten eher für eine variablenorientierte und Distanzmaße für eine ojektorientierte Clusteranalyse verwendet werden (vgl. Bacher 1996, S. 198f.).

Definition 2. Unähnlichkeits- oder Distanzmaß

Sei $X = \{x_1, \ldots, x_n\}$ eine endliche Menge. Eine Abbildung d: $X \times X \rightarrow \mathbb{R}$ heißt Abstands- bzw. Distanzfunktion, wenn für zwei Objekte $x_i, x_j \in X$ gilt:

1. $d(x_i, x_j) = d(x_j, x_i)$ $\quad x_i, x_j \in X$
2. $d(x_i, x_j) \geq 0$ $\quad x_i, x_j \in X$
3. $d(x_i, x_i) = 0$ $\quad x_i \in X$

Eine Abstandsfunktion heißt metrisch, wenn zusätzlich zwei weitere Eigenschaften gelten (vgl. Späth 1977, S. 14f.):

4. Aus $d(x_i, x_j) = 0$ für $x_i, x_j \in X$ folgt die Identität der Objekte $x_i = x_j$.
5. $d(x_i, x_j) \leq d(x_i, x_k) + d(x_k, x_j)$ für $x_i, x_j, x_k \in X$ (Dreiecksungleichung)

Eine Matrix, deren Einträge aus den Distanzen[2] bzw. Unähnlichkeitsmaßen $d_{ij} = d(x_i,x_j)$ bestehen, wird als Distanzmatrix von X bezeichnet.

Das bekannteste Distanzmaß ist die Euklidische Distanz, die von der L_2-Norm eines Vektors $x \in \mathbb{R}^n$

$$\|x\|_2 = \sqrt{\sum_{k=1}^{n} x_k^2} = \sqrt{x^T x}$$

hergeleitet wird zu

$$d_2(x_i, x_j) = \| x_i - x_j \|_2 = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \qquad x_i, x_j \in \mathbb{R}^n.$$

Dieses Maß ist invariant bzgl. der Translation und der orthogonalen linearen Transformation (vgl. Späth 1977, S. 16). Es eignet sich für intervallskalierte und unkorrelierte Variablen (vgl. Kaufman/Rousseeuw 1990, S. 11f.). Eine Verallgemeinerung der Euklidischen Abstandsfunktion ist die L_p-Distanz (oder Minkowski-Distanz[3])

$$d_p(x_i, x_j) = \| x_i - x_j \|_p = \left(\sum_{k=1}^{n} |x_{ik} - x_{jk}|^p \right)^{1/p} \qquad x_i, x_j \in \mathbb{R}^n,\ p \in \mathbb{R},\ p \geq 1.$$

Auch dieses Maß ist invariant bzgl. der Translation. Für $p = 2$ erhält man das Euklidische Distanzmaß und für $p = 1$ die Manhattan Distanz (oder City Block Distanz)

$$d_1(x_i, x_j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{in} - x_{jn}| \qquad x_i, x_j \in \mathbb{R}^n.$$

Bei dieser Abstandsfunktion geht man von der Überlegung aus, dass man in einer Stadt mit geradlinigem Straßennetz, d.h. mit vertikalen bzw. horizontalen Straßen, nur über ein (rechtwinkliges) Dreieck vom Ausgangspunkt i zum Zielpunkt j kommt (vgl. Kaufman/ Rousseeuw 1990, S. 12f). Ein weiterer Spezialfall der L_p-Distanz ergibt sich durch die Wahl des Parameters $p = \infty$:

$$d_\infty(x_i, x_j) = \| x_i - x_j \|_\infty = \max_{1 \leq k \leq n} |x_{ik} - x_{jk}|.$$

Die Mahalanobis-Distanz basiert auf einer anderen Verallgemeinerung der Euklidischen Distanz, und zwar geht man von der Norm

---

[2] In der vorliegenden Arbeit werden die Begriffe Distanz und Unähnlichkeit synonym verwendet.

[3] Die L_p-Distanz ergibt sich aus der verallgemeinerten Minkowski-Distanz $d(x_i, x_j) = \left(\sum_{k=1}^{n} |x_{ik} - x_{kj}|^p \right)^{1/q}$ mit $p, q \in \mathbb{R}$, $p,q \geq 1$ und $p = q$.

$$\|x\|_B = \sqrt{x^T B x} \text{ mit } x^T B\, x \geq 0 \qquad \forall\, x \in \mathbb{R}^n \text{ und } x^T B\, x = 0 \Leftrightarrow x = 0.$$

aus und erhält die entsprechende Metrik der Form

$$d_B(x_i, x_j) = \sqrt{(x_i - x_j)^T \cdot B \cdot (x_i - x_j)}.$$

Die Matrix *B* ist bei dieser Metrik die Inverse der Kovarianzmatrix der Variablen und man bezieht sich bei der Berechnung auf die entsprechende Spalte bzw. Zeile. Die Mahalanobis-Distanz besitzt die allgemeinste Invarianzeigenschaft, d.h. sie ist invariant bzgl. nichtsingulärer Transformationen *C* (vgl. Späth 1977, S. 17f.). Das bedeutet, dass die Daten so transformiert werden, dass sie standardisiert und unkorreliert sind und somit die Maßeinheiten der Variablen keine Bedeutung haben. Der Nachteil liegt jedoch darin, dass sich die Metrik auf alle Objekte gemeinsam auswirkt und die Berechnung um einiges aufwändiger ist (vgl. Späth 1977, S. 19; vgl. Pruscha 2006, S. 292).

Bei der variablenorientierten Clusteranalyse werden Korrelationskoeffizienten verwendet (vgl. Kaufman/Rousseeuw 1990, S. 16-20). Dabei können sowohl der Pearson-Produkt-Moment-Korrelationskoeffizient (für die Variablen *f* und *g)*

$$R_P(f,g) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_{if} - \bar{x}_f)(x_{ig} - \bar{x}_g)}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_{if} - \bar{x}_f)^2 \frac{1}{n}\sum_{i=1}^{n}(x_{if} - \bar{x}_g)^2\right)^{1/2}}$$

als auch der Spearman-Korrelationskoeffizient

$$R_S(f,g) = \frac{\sum_{i=1}^{n}(r_i^f - \bar{r})(r_i^g - \bar{r})}{\sum_{i=1}^{n}(r_i^f - \bar{r})^2 \sum_{i=1}^{n}(r_i^g - \bar{r})^2}$$

mit r_i^f, r_i^g Rangzahlen der entsprechenden Variablen und $\bar{r} = \frac{1}{2}(n+1)$

benutzt werden. Die Werte beider Koeffizienten liegen zwischen *–1* und *+1*, wobei ein kleiner Wert eine große Unähnlichkeit nahe legt. Damit der Korrelationskoeffizient einem Distanzmaß entspricht, ist eine der beiden Transformationen

1. $d(x_i, x_j) = \frac{1 - R(x_i, x_j)}{2}$ $\qquad x_i, x_j \in \mathbb{R}^n$

2. $d(x_i, x_j) = 1 - |R(x_i, x_j)|$ $\qquad x_i, x_j \in \mathbb{R}^n$

durchzuführen (vgl. Kaufman/Rousseeuw 1990, S. 19f.). Bei der ersten Transformation bekommen zwei Variablen mit einer hohen positiven Korrelation und bei einer starken negativen Korrelation einen Wert nahe null, also eine Einstufung als sehr unähnlich. Bei

der zweiten Transformation bedeutet eine stark negative oder hohe positive Korrelation, dass die Variablen sich ähneln (vgl. Kaufman/Rousseeuw 1990, S. 19).

Bei der Clusteranalyse kann man natürlich auch mit Ähnlichkeitsmaßen statt mit Distanzen arbeiten.

Definition 3. Ähnlichkeitsmaß

Sei $X = \{x_1, \ldots, x_n\}$ eine endliche Menge. Eine Abbildung $s: X \times X \rightarrow \mathbb{R}$ heißt Ähnlichkeitsfunktion, wenn für zwei Objekte $x_i, x_j \in X$ gilt:

1. $s(x_i, x_j) = s(x_j, x_i)$ $\quad x_i, x_j \in X$

2. $s(x_i, x_j) \geq 0$ $\quad x_i, x_j \in X$

2. $s(x_i, x_j) \leq s(x_i, x_i)$ $\quad x_i, x_j \in X$

3. $s(x_i, x_i) = 1$ $\quad x_i, \in X$

(vgl. Kaufman/Rousseeuw 1990, S. 20).

Je größer der Wert der Ähnlichkeitsfunktion zweier Objekte ist, desto ähnlicher sind sie. Die Werte $s_{ij} = s(x_i, x_j)$ können in einer Matrix, der sogenannten Ähnlichkeitsmatrix, eingetragen werden.

Für eine variablenorientierte Clusteranalyse kann der Korrelationskoeffizient nach Pearson oder nach Spearman verwendet werden, wobei hier von *R*-Korrelation gesprochen wird (vgl. Kaufman/Rousseeuw 1990, S. 20f.; vgl. Bacher 1996, S. 221). Da diese Maße negative Werte annehmen können, sollten sie transformiert werden. Kaufman und Rousseeuw (1990) schlagen zwei Transformationen für das (allgemeine) Korrelationsmaß *R* vor (vgl. Kaufman/Rousseeuw 1990, S. 21):

1. $s(x_i, x_j) = \dfrac{1 + R(x_i, x_j)}{2}$ $\quad x_i, x_j \in \mathbb{R}^n$

2. $s(x_i, x_j) = |R(x_i, x_j)|$ $\quad x_i, x_j \in \mathbb{R}^n$

Wie bei den Unähnlichkeitsmaßen stuft die erste Transformation Objekte, die gegensätzlich orientiert sind, d.h. $R(x_i, x_j) = -1$, als unähnlich ein ($s(x_i, x_j) = 0$), wobei die zweite Transformation von der Überlegung ausgeht, dass die beiden Variablen sehr wohl einen hohen Zusammenhang haben und zu einer Gruppe zusammengefasst werden sollten ($R(x_i, x_j) = \pm 1 \Rightarrow s(x_i, x_j) = 1$). Einige Methoden der Clusteranalyse verwenden nur Unähnlichkeitsmaße, darum kann mithilfe der Transformation

$$d(x_i, x_j) = 1 - s(x_i, x_j) \quad x_i, x_j \in \mathbb{R}^n$$

aus den Werten für Ähnlichkeiten die entsprechenden Distanzmaße berechnet werden (vgl. Kaufman/Rousseeuw 1990, S. 21f.).

1.3 Fehlende Werte

Für die Clusteranalyse ist es wesentlich, dass die Distanzen ermittelt werden können. Daher ist es notwendig, sich zu überlegen, wie fehlende Werte in der Datenmatrix behandelt werden sollen. Trivialerweise werden Objekte bzw. Merkmale, bei denen keine Messdaten aufscheinen, gestrichen, da sie für die weitere Untersuchung keine Erkenntnisse liefern. Für den Fall, dass bei zwei Objekten nur Werte für unterschiedliche Variablen vorliegen, können die Distanzen nicht direkt berechnet werden. Um dennoch zu einer gültigen Distanzmatrix zu gelangen, kann man eine der folgenden Aktionen wählen (vgl. Kaufman/Rousseeuw 1990, S. 14f.; vgl. Bacher 1996, S. 230ff.):

1. Fallweises Ausscheiden: Man streicht die Objekte, bei denen Daten fehlen und berechnet die Distanzmatrix für die Restdaten. Das kann jedoch dazu führen, dass die Anzahl der Fälle bzw. Merkmale erheblich verringert wird.
2. Paarweises Ausschließen: Dabei werden die beiden Variablen bzw. die beiden Fälle gestrichen und es wird mit der Restmatrix weiter gearbeitet. Damit verringert sich zwar die Anzahl der Fälle bzw. Variablen, aber nicht in dem Ausmaß wie bei der ersten Methode.
3. Als Alternative können durchschnittliche Distanzwerte auf Basis der vorhandenen Daten berechnet werden, die die fehlenden Werte ersetzen. Somit hat man die gesamte Anzahl an Objekten und Merkmalen in der Untersuchung, jedoch ist der Fehler, der durch die Substitution der Durchschnittswerte entsteht, bei der Interpretation der Ergebnisse zu berücksichtigen.
4. Es wird die ursprüngliche Datenmatrix korrigiert, indem fehlende Messdaten durch Mittelwerte ersetzt werden.

Bacher (1996) hat in einer Untersuchung über die Auswirkungen der Methoden zur Behandlung fehlender Werte auf eine fehlerhafte Klassifizierung festgestellt, dass der Einfluss relativ gering ist und sich die Ergebnisse der einzelnen Methoden sich wiederum nur geringfügig voneinander unterscheiden (vgl. Bacher 1996, S. 230f.).

2 Einige Clusteranalyseverfahren

Im Folgenden werden exemplarisch einige Clusteranalyseverfahren etwas genauer beschrieben.

1. Nächste-Nachbar-Verfahren (nearest neighborhood), insbesondere *complete-linkage*
2. Repräsentantenverfahren (*k-medoid*-Verfahren, *PAM*)
3. Verfahren zur Konstruktion von Clusterzentren (*k-means*-Verfahren)

Diese Verfahren werden häufig verwendet und sind auch in der Statistik-Software *R* implementiert. Sie stellen sozusagen die „Prototypen" der Clusteranalyseverfahren dar.

2.1 Nächste-Nachbarn-Verfahren (nearest neighborhood)

Nächste-Nachbarn-Verfahren sind hierarchisch agglomerative Verfahren und werden zu den deterministischen Methoden gezählt (vgl. Bacher 1996, S. 142f.). Mit dem Begriff „nächster Nachbar" zu einem Klassifikationsobjekt x ist jenes Objekt gemeint, dessen Wert des Ähnlichkeitsmaßes (oder Unähnlichkeitsmaßes) größer/gleich (oder kleiner/ gleich) einem festgelegten Schwellenwert ist. Man kann nun fordern, dass in einem Cluster eine bestimmte Anzahl an nächsten Nachbarn vorhanden sein soll oder jedes Objekt einer Gruppe mindestens einen *b*-ten Nachbar hat. Auf diesem Prinzip beruhen sehr viele Methoden. In einigen Übersichtsarbeiten werden bis zu 40 unterschiedliche Verfahren aufgelistet (vgl. Bacher 1996, S. 142).

2.1.1 Grundalgorithmus der hierarchisch agglomerativen Verfahren

1. *Initialisierung*

   $k := n$

   Startpartition $\mathcal{A}^{(0)} = \{A_1,\ldots,A_n\}$ mit $A_i = \{x_i\}$ für $i = 1,\ldots,n$,

   d.h. jedes Klassifikationsobjekt bildet ein selbständiges Cluster.

2. *Verschmelzung zweier Cluster*

   Suchen nach zwei Clustern A_i und A_j mit $i \neq j$, die die größte Ähnlichkeit (bzw. die geringste Unähnlichkeit) aufweisen, und Verschmelzung der beiden zu einem neuen Cluster $A = A_i \cup A_j$.

   $k := k - 1$

3. *Berechnung des Ähnlichkeits- bzw. Unähnlichkeitsmaßes*
   Solange *k > 1*
   Berechnung des Ähnlichkeits- bzw. Unähnlichkeitsmaßes des neu gebildeten Clusters *A* zu den verbleibenden Clustern *B*.
4. *Iteration*
   Fortsetzung bei Schritt 2.

Im Schritt 3 bricht der Algorithmus ab, wenn die gröbste Partition erreicht ist, und zwar jene mit nur einem Cluster für alle Klassifikationsobjekten (*k = 1*). Nach diesem Grundalgorithmus gehen das Nächste-Nachbarn-Verfahren, die Mittelwerteverfahren und einige Verfahren zur Konstruktion von Clusterzentren (*Median-*, *Zentroid-* und *Ward-*Verfahren) vor. Der Unterschied liegt im Schritt 3, bei der Berechnung der Ähnlichkeiten bzw. Unähnlichkeiten der neu gebildeten Gruppe zu allen übrigen Gruppen (vgl. Bacher 1996, S. 144; vgl. Pruscha 2006, S. 288f.).

2.1.2 *Complete-linkage* bzw. *single-linkage* als Basismodell

Complete- und *single-linkage* kann man als Basismodelle für hierarchisch agglomerative Verfahren betrachten. Beide Verfahren können mit sehr allgemeinen Daten arbeiten, d.h. sie benötigen wenige Voraussetzungen (vgl. Bacher 1996, S. 145). Bei beiden Verfahren kann eine direkt eingegebene Distanz- oder Ähnlichkeitsmatrix verwendet werden. Diese Verwendung ist nur für nicht-metrische Daten geeignet, da nur die ordinale Information der Ähnlichkeiten bzw. der Unähnlichkeiten genutzt wird (vgl. Bacher 1996, S. 145f.).

Beim *complete-linkage*-Verfahren (*Methode des weitest entfernten Nachbarn* oder *Maximum-Methode*) steht die Forderung der Homogenität innerhalb eines Clusters im Vordergrund. Das bedeutet, dass alle Objekte innerhalb einer Gruppe zueinander die nächsten Nachbarn sind, daher werden die Cluster auch Cliquen genannt (vgl. Bacher 1996, S. 239f.). Beim *single-linkage*-Verfahren (*Methode des nächsten Nachbarn* oder *Minimum-Methode*) ist diese Homogenitätsforderung abgeschwächt; nun soll zumindest jedes Element eines Clusters einen nächsten Nachbarn haben.

Beim *complete-linkage* wird im Schritt 3 des Grundalgorithmus das neue Maß dem größeren der beiden alten Ähnlichkeiten (bzw. dem kleineren der alten Unähnlichkeit)

$$s_{A,B} = \max\left(d_{A_i,B}, d_{A_j,B}\right) \text{ bzw. } d_{A,B} = \min\left(d_{A_i,B}, d_{A_j,B}\right)$$

und beim *single-linkage* das neue Ähnlichkeitsmaß dem kleineren der beiden alten Ähnlichkeiten (bzw. die neue Distanz dem größeren der alten Distanzen)

$$s_{A,B} = \min\left(d_{A_i,B}, d_{A_j,B}\right) \text{ bzw. } d_{A,B} = \max\left(d_{A_i,B}, d_{A_j,B}\right)$$

gleichgesetzt (vgl. Pruscha 2006, S. 298). Hierarchische Verfahren lassen sich gut graphisch darstellen, und zwar in Form von baumartigen Diagrammen, den sogenannten Dendogrammen. Dabei gibt jede Verzweigung die Partition für jedes *k* wieder.

Beide Verfahren weisen einige Schwachstellen auf. So kann es beim *single-linkage* vorkommen, dass sich unterscheidende Cluster aufgrund der schwachen Homogenitätsbedingung verschmolzen werden (Kontraktionseffekt). Dagegen kann es beim *complete-linkage* passieren, dass zu viele Cluster mit wenigen Objekten gebildet werden (Dilatationseffekt). Zudem eignen sich diese beiden Methoden eher für eine geringe Anzahl an Klassifikationsobjekten, da die Un- bzw. Ähnlichkeitsmatrix im Arbeitsspeicher mitgeführt werden muss. Um diese Nachteile auszumerzen, wurden Modifikationen der Algorithmen gebildet und somit andere Verfahren entwickelt (vgl. Bacher 1996, S. 145f.).

2.2 Verfahren zur Konstruktion von Clusterzentren

Bei der partitionierenden Methode wird vorab festgelegt, in wie viele Cluster die Daten eingeteilt werden. Der Algorithmus gibt daraufhin eine Lösung mit der Aufteilung der *n* Objekte in *k* Gruppen zurück. Daher ist es notwendig, sich mit der optimalen Klassenanzahl auseinander zu setzen. Vielfach geschieht dies derart, dass der Algorithmus für verschiedene *k* durchlaufen wird und z. B. mithilfe von Screeplot oder Silhouette-Plot eine Entscheidung getroffen wird (vgl. Kaufman/Rousseeuw 1990, S. 38f.).

2.2.1 *k*-means Verfahren

Das *k-means* Verfahren ist algorithmisch gesehen eine nicht-hierarchische Clusteranalysemethode. Zudem zählt es zu den Verfahren zur Konstruktion von Clusterzentren, bei denen die *k* repräsentativen Objekte mithilfe der Mittelwerte der Cluster berechnet werden und die übrigen Objekte aufgrund ihrer „Ähnlichkeit" zu diesen Clusterzentren – der Euklidischen Distanz – den einzelnen Gruppen zugeordnet werden. Strenggenommen darf diese Methode nur bei quantitativen Variablen angewandt werden, doch durch einige „Tricks" wie durch Einführen von Dummy-Variablen wird das *k*-means-Verfahren auch für ordinale, dichotome oder nominale Daten verwendet (vgl. Bacher 1996, S. 297).

2.2.2 Grundalgorithmus[4]

1. *Initialisierung*

   Eingabe (oder Berechnung) von *k* Clusterzentren, damit erhält man eine Startpartition $\mathcal{A}^{(0)}$

2. *Neuberechnung der Clusterzentren*

   Sei $\mathcal{A}^{(m)}$ eine Partition.

   Berechnung des Gruppenschwerpunktes $\bar{x}_A = \frac{1}{n_A} \sum_{x_j \in A} x_{ij}$ für jedes Cluster $A \in \mathcal{A}^{(m)}$

   (n_A ... Anzahl der Objekte in Cluster *A* mit gültigen Werten in den Variablen *j*)

3. *Zuordnung der Objekte*

   Verschieben der *n* Objekte in diejenige Gruppe, deren Schwerpunkt dem Objekt am nächsten liegt (Euklidischen Distanz)

   Partition $\mathcal{A}^{(m+1)}$

4. *Iteration*

   Fortsetzung bei Schritt 2.

   Wenn im Schritt 3 kein Klassifizierungsobjekt die Gruppe wechselt, endet das Verfahren.

Wenn das Verfahren wie vorgesehen mit einer Partition $\mathcal{A}^* = (A_1, \ldots, A_n)$ endet, so hat man mit den Gruppenschwerpunkten

$$a_1^* = \bar{x}_{A_1}, \ldots, a_k^* = \bar{x}_{A_k}$$

die Clusterzentren für die Beobachtungsvektoren $x_1, \ldots, x_n$ gefunden (vgl. Bacher 1996, S. 306f.).

2.2.3 Modifikationen des Grundalgorithmus

Durch Veränderung der einzelnen Schritte des oben beschriebenen Algorithmus ergeben sich die Varianten des *k-means*-Verfahrens. Diese betreffen die Berechnung der Startwerte (Clusterzentren), die verwendete Zuordnungs- oder Distanzfunktion und die Anforderungen an die zu berechnenden Clusterzentren (vgl. Bacher 1996, S. 228f.).

[4] Eine detailliertere Beschreibung des Grundalgorithmus findet man bei Bacher (vgl. Bacher 1996, S. 280) und bei Pruscha (Pruscha 2006, S. 307f.).

Bei der Auswahl der Startwerte gibt es folgende Möglichkeiten:

1. Man wählt die ersten *k* Objekte als Clusterzentren.
2. Die Klassifizierungsobjekte werden zufällig den *k* Clustern zugeordnet.
3. Die Startwerte werden mit einem anderen Clusteranalyseverfahren berechnet, z. B. durch ein Repräsentanten-Verfahren.
4. Der/Die Anwender/in bestimmt Startwerte, die bereits eine Struktur aufgrund von inhaltlichen Überlegungen vorgeben.

Eine weitere Modifikation besteht darin, dass nach jeder Neuzuordnung eines Objektes die Clusterzentren nochmals berechnet werden. Bei der Methode „*hill-climbing*" wird ein Objekt nur dann neu zugeordnet, wenn sich dadurch der Wert der Minimierungsfunktion ändert (vgl. Bacher 1996, S. 344f.).

Eine andere Form der Veränderung ergibt sich, wenn man das Distanzmaß ändert. Bekannt sind Modifikationen mit der Mahalanobis-Distanz, wobei die empirische, gepoolte Kovarianzmatrix oder eine Diagonalmatrix verwendet werden. Der Vorteil liegt darin, dass die Varianzen und Korrelationen zwischen den Variablen beseitigt werden (vgl. Bacher 1996, S. 346).

2.3 Repräsentantenverfahren

Bei den Repräsentantenverfahren möchte man *k* „typische" Objekte in der Menge der Klassifizierungsobjekte finden, die die Struktur der Daten so gut wie möglich widerspiegeln. Diese repräsentativen Objekte werden auch als „centrotypes", „medoids" oder „leading case" bezeichnet (vgl. Bacher 1996, S. 279; vgl. Kaufman/Rousseeuw 1990, S. 69). Die restlichen Objekte werden danach dem „nächsten" repräsentativen Objekt zugeordnet, wodurch die *k* Cluster entstehen. Entscheidend ist die Wahl der repräsentativen Elemente, da sie das Zentrum der Cluster darstellen. Das geschieht derart, dass die durchschnittliche Distanz (oder average dissimilarity) dieser Medoids zu allen anderen Objekten minimiert wird (vgl. Kaufman/Rousseeuw 1990, S. 40f.).

Das Finden der repräsentativen Objekte ist der große Vorteil der Repräsentantenmethode. Diese ermöglichen eine Charakterisierung der Cluster und können für die Interpretation der Ergebnisse und für weitere Analysen verwendet werden. Zusätzlich bietet dieses Verfahren eine graphische Darstellung, den sogenannten Silhouette-Plot (vgl. Kaufman/Rousseeuw 1990, S. 71).

2.3.1 Grundalgorithmus[5]

1. *Initialisierung*

   Berechnung der Zahl der nächsten Nachbarn (nn_i) für jedes Objekt *i*

   Berechnung der Distanz d_{ij} des Objektes *i* zu Objekt *j*

   S_{homo} … Schwellenwert für die Homogenität innerhalb der Cluster

   S_{hetero} … Schwellenwert für die Heterogenität zwischen den Cluster

2. *Ordnen der Objekte nach der Größe* nn_i

3. *Bestimmung der Repräsentanten (hierarchisch)*

   Objekt *i* wird Repräsentant, wenn die Distanzen zu anderen Repräsentanten *j* größer als der Schwellenwert für die Heterogenität ist, $d_{ij} > S_{hetero}$

4. *Zuordnung der Objekte*

   Objekt *j* kommt zu Cluster *i*, wenn die Distanz zum Repräsentanten *i* kleiner als der Schwellenwert für die Homogenität ist, $d_{ij} < S_{homo}$

2.3.2 *k-medoid*-Method oder *k-median*: *Partioning Around Medoids*

Beim *Partioning Around Medoids*-Verfahren, kurz *PAM*, werden Cluster gesucht, die nahezu kugelförmig sind. Diese Methode ist daher nicht geeignet, Strukturen mit anderen Formen zu finden. Bei diesem Verfahren ist der mittlere Abstand aller Objekte innerhalb eines Clusters zum Repräsentanten minimal (vgl. Kaufman/Rousseeuw 1990, S. 71f.).

2.4 Graphische Darstellung von partitionierenden Methoden

Bei hierarchischen Methoden bietet das Dendogramm eine gute Darstellung der Cluster und ihrer Elemente. Das Ergebnis bei partitionierenden Methoden dagegen besteht aus einer Liste der Cluster mit den entsprechenden Objekten. Ein Vergleichen bzw. eine Interpretation der verschiedenen Partitionen ist damit nicht immer einfach. Eine Möglichkeit, um Unterschiede bzw. Gemeinsamkeiten auf einen ersten Blick zu erkennen, bietet der Silhouette-Plot, bei der Kennzahlen *s*(*i*) für die Qualität der Zuordnung der Objekte als Balken dargestellt werden (vgl. Rousseeuw 1987, S. 54f).

[5] Eine detaillierte Beschreibung mit ausführlichen Beispielen findet sich bei Bacher (vgl. Bacher 1996, S. 281-284).

Zur Berechnung dieser Kennzahlen *s*(*i*) wird für jedes Objekt *i* die durchschnittliche Distanz *a*(*i*) zu allen anderen Objekten, die sich in demselben Cluster befinden, ermittelt. Daraufhin wird jenes Cluster gesucht, das dem Objekt *i* am nächsten liegt, indem die durchschnittliche Distanz (*d*(*i*,*C*)) von *i* zu allen Elementen in den Clustern *C* berechnet und davon jene mit der geringsten Entfernung gewählt wird

$$b(i) = \min_{i \notin C} d(i, C).$$

Dieses Cluster wird auch „Nachbar" von Objekt *i* genannt. Mithilfe der Formel

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

ergibt sich nun die Kennzahl *s*(*i*) für jedes Objekt *i*, die auch als eine Art von Unähnlichkeit zwischen *i* und seinem Nachbar-Cluster betrachtet werden kann(vgl. Rousseeuw 1987, S. 55f.). Wenn ein Cluster nur aus einem Element besteht, wird *s*(*i*) gleich null gesetzt.

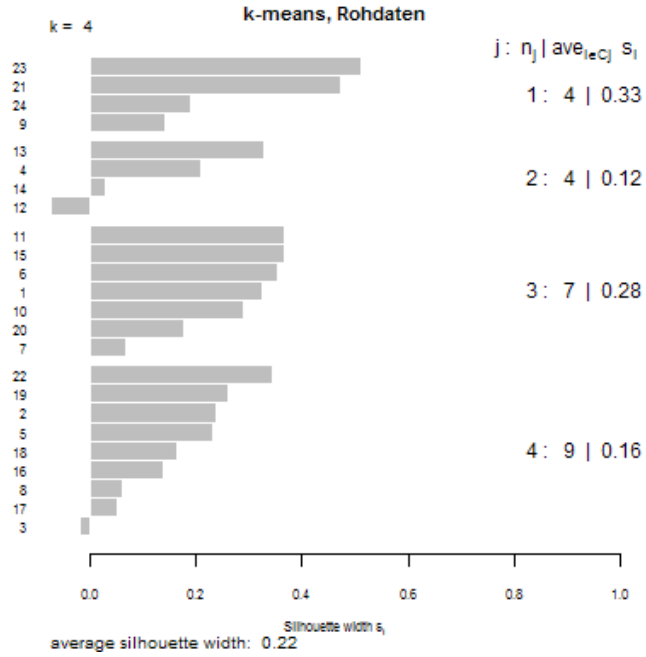


Abb. 3.2.1. Silhouette-Plot

Der Wert dieser Kennzahl liegt zwischen *–1* und *+1*. Hat ein Objekt einen Wert nahe eins, dann ist das gleichbedeutend damit, dass es gut in dieses Cluster passt. Bei Werten nahe null ist die minimale Distanz *b*(*i*) von *i* zu den anderen Clustern in der Größenordnung der durchschnittlichen Distanz innerhalb des Clusters. Damit liegt das Objekt zwischen zwei Cluster. Bei negativen Werten ist das Objekt anscheinend nicht gut zugeordnet, es scheint der benachbarten Gruppe sehr nahe zu sein.

Für den Silhouette-Plot werden nun die Kennzahlen *s*(*i*) als Balken dargestellt, wobei die Objekte nach den Clustern sortiert sind und man somit auch die Gruppengröße erkennen kann. Zusätzlich werden die durchschnittlichen Silhouette-Breiten eines Clusters aus den Kennzahlen *s*(*i*) der Clusterelemente berechnet. Der Durchschnitt über alle Kennzahlen *s*(*i*) ergibt die durchschnittliche Silhouette-Breite dieser Partition (vgl. Rousseeuw 1987, S. 56-59).

Abbildung 3.2.1 zeigt einen Silhouette-Plot für eine Clusterung mit vier Klassen. Man erhält die Information, dass sich im ersten und zweiten Cluster jeweils vier Objekte, im

dritten sieben und im letzten neun Elemente befinden (rechte Seite des Plots). Die durchschnittliche Silhouette-Breite ist mit 0.22 recht niedrig, daher sollte man auf alle Fälle die durchschnittlichen Silhouette-Breiten der einzelnen Cluster betrachten. Man sieht sofort, dass zwei Cluster eine niedrige Silhouette-Breite haben (0.12 bzw. 0.16), d.h. die Kennzahlen $s(i)$ innerhalb dieser Gruppen dürften eher gering sein oder sogar negativ.

Die Kennzahlen $s(i)$ (Silhouette-Breiten) können außerdem ein Hilfsmittel sein, um eine geeignete Anzahl an Klassen zu ermitteln. Dafür werden die durchschnittlichen Silhouette-Breiten für Partitionen mit unterschiedlicher Klassenanzahl k verglichen. Jene mit dem höchsten Wert scheint die „beste" Aufteilung der Objekte zu haben, wobei man auch die Cluster-Silhouette-Breiten und die Kennzahlen $s(i)$ beachten soll. Man bedenke den Fall, dass ein Cluster nur aus einem Objekt besteht, d.h. die durchschnittliche Silhouette-Breite wird dann eher gering sein und trotzdem kann die Partition die Struktur der Daten gut widerspiegeln.

IV

EMPIRISCHE AUSWERTUNG

1 Daten und Vorgehensweise

Das Land Steiermark stellt Daten zur Luftqualität mithilfe des Landes-Umwelt-Informationssystem (LUIS) zur Verfügung (http://www.umwelt.steiermark.at). Die Messwerte liegen halbstündlich vor und können als Rohdaten oder bereits aggregiert (Tages-, Monats- bzw. Jahresmittelwerte) herunter geladen werden. Für die vorliegende Untersuchung wurden 25 Messstationen und die 4 Merkmale Feinstaub PM_{10} [µg/m³], Ozon O_3 [µg/m³], Schwefeldioxid SO_2 [µg/m³] und Stickstoffdioxid NO_2 [µg/m³] ausgewählt. Der Untersuchungszeitraum der vorliegenden Arbeit erstreckt sich über drei Jahre (= 1096 Tage), und zwar vom 1. 1. 2008 bis 31. 12. 2010, wobei die Daten als Tagesmittelwerte aufbereitet wurden.

Nicht alle Messstationen erheben alle vier Merkmale. Die Luftschadstoffe Feinstaub PM_{10} und Stickstoffdioxid NO_2 werden bei allen 25 Messstellen erhoben, Ozon O_3 von 10 Stationen und Schwefeldioxid SO_2 von 19 (siehe Tabelle 4.1). Zudem gibt es bei den meisten Stationen nicht für jeden Tag und über den gesamten Zeitraum Daten, sondern es fehlen Werte. Manchmal sind diese Lücken nur einige Tage lang, in einigen Fällen gibt es Ausfälle, die einige Wochen bzw. Monate dauern. Für jeden Luftschadstoff L liegen somit diskrete Messwerte y_{ij} vor, wobei die Indices $i = 1, \ldots, n_L$ für die Messstationen und $j = 1, \ldots, 1096$ für die Tage der Messung stehen. In Tabelle 4.1 findet man eine Übersicht über die Anzahl der Messwerte (2. Spalte) und fehlende Werte (1. Spalte) pro Messstation.

Es gibt verschiedene Vorgehensweisen für die Behandlung von fehlenden Daten. Eine besteht darin, dass die Daten aggregiert werden, z. B. als Wochen- oder Monatsmittelwerte. Der Nachteil dieser Methode liegt im Verlust von Informationen, der bei der Inter-

pretation der Ergebnisse berücksichtigt werden muss. Zudem eignet sich diese Möglichkeit nicht für die Daten der Messstation Graz Mitte Gries, da hier für annähernd zwei Jahre (2008 – 2009) Messwerte fehlen. Eine weitere Option besteht darin, Objekte mit fehlenden Werten auszuschließen. Für die vorliegende Arbeit werden nun jene Messstationen ausgeschlossen, für die mehr als die Hälfte der Daten fehlen. Bei den Merkmalen Feinstaub PM_{10} und Stickstoffdioxid NO_2 wird die Messstelle Graz Mitte Gries und für Schwefeldioxid SO_2 werden Kapfenberg und Leibnitz ausgeschlossen.

i	*Messstation*	*PM_{10} [µg/m³]*		*O_3 [µg/m³]*		*SO_2 [µg/m³]*		*NO_2 [µg/m³]*	
	n_L	25		10		18		25	
		fehlend	*vor-handen*	fehlend	*vor-handen*	fehlend	*vor-handen*	fehlend	*vor-handen*
1	Bruck a. d. Mur	17	1079	—	—	20	1076	9	1087
2	Deutschlandsberg AK	16	1080	7	1089	10	1086	18	1078
3	Fürstenfeld	3	1093	2	1094	2	1094	21	1075
4	Judenburg	3	1093	12	1084	—	—	18	1078
5	Judendorf-Süd	47	1049	—	—	22	1074	17	1079
6	Kapfenberg	3	1093	—	—	1026	70	213	883
7	Knittelfeld Parkstraße	1	1095	—	—	0	1096	1	1095
8	Köflach	11	1085	—	—	5	1091	13	1083
9	Leibnitz	52	1044	—	—	974	122	3	1093
10	Leoben Donawitz	47	1049	—	—	46	1050	39	1057
11	Leoben Göß	92	1004	—	—	333	763	13	1083
12	Liezen	14	1082	26	1070	7	1089	7	1089
13	Masenberg	29	1067	20	1076	23	1073	63	1033
14	Mürzzuschlag	128	968	6	1090	—	—	8	1088
15	Niklasdorf	29	1067	—	—	7	1089	11	1085
16	Peggau	15	1081	—	—	363	733	2	1094
17	Straßengel Kirche	103	993	—	—	38	1058	25	1071
18	Voitsberg	170	926	19	1077	3	1093	25	1071
19	Weiz	18	1078	2	1094	—	—	26	1070
20	Zeltweg Hauptschule	0	1096	—	—	—	—	54	1042
21	Graz Don Bosco	25	1071	—	—	—	—	25	1071
22	Graz Mitte Gries	683	413	—	—	—	—	702	394
23	Graz Nord	27	1069	25	1071	46	1050	20	1076
24	Graz Süd Tiergartenweg	0	1096	9	1087	31	1065	2	1094
25	Graz West	38	1058	—	—	312	784	9	1087
	gesamt		25829		10832		17556		26056

Tab. 4.1. Fehlende Werte und Anzahl der Messwerte y_{ij} pro Luftschadstoff und Messstation

Bei der Betrachtung der Zeitreihenplots für die Luftschadstoffe Feinstaub, Ozon, Schwefeldioxid und Stickstoffdioxid (über alle Messstationen hinweg) kann man einige Besonderheiten erkennen (siehe Abbildung 4.1.1).

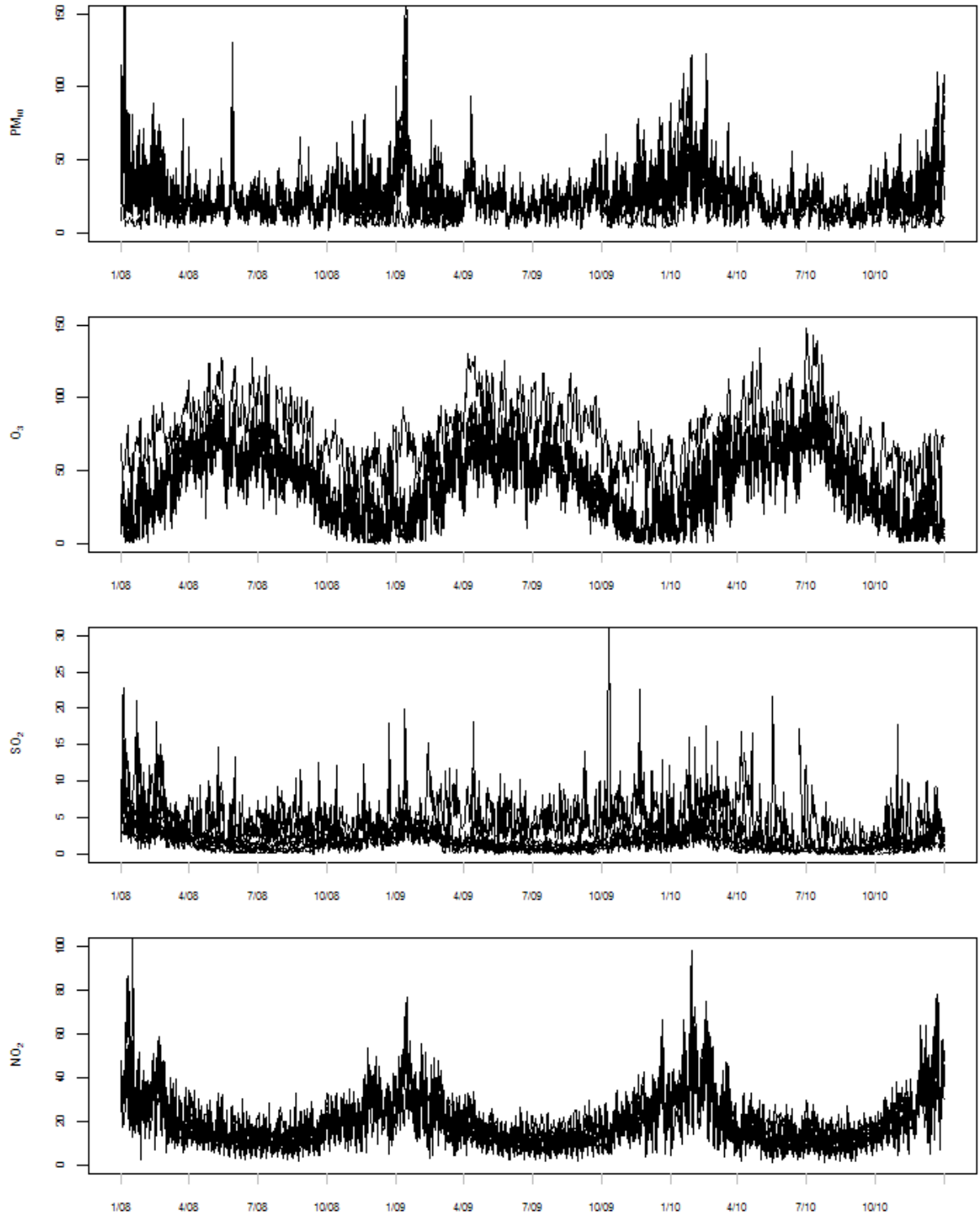


Abb. 4.1.1. Zeitreihen der Luftschadstoffe über alle n_L Messstationen: Feinstaub PM_{10} (1. Plot) bei 25 Stationen, Ozon O_3 (2. Plot) bei 10 Stationen, Schwefeldioxid SO_2 (3. Plot) bei 19 Stationen, Stickstoffdioxid NO_2 (4. Plot) bei 25 Stationen

Feinstaub kommt einerseits auf sehr niedrigem Konzentrationsniveau vor, da es über den Untersuchungszeitraum auch Werte nahe null gibt, andererseits existieren stark ausgeprägte Spitzen, die erwartungsgemäß um Neujahr zu finden sind. Zudem sind die Werte in den Wintermonaten (Oktober – März) höher als in den Sommermonaten (April – September).

Bei den Luftschadstoffen Ozon und Stickstoffdioxid sind saisonale Schwankungen an allen Messstellen erkennbar, mit stark ausgeprägten Maxima zu ähnlichen Zeitpunkten: bei Ozon in den Sommermonaten (Juni, Juli, August) und beim Stickstoffdioxid in den Wintermonaten (Jänner, Februar). Schwefeldioxid kommt generell in niedrigeren Konzentrationen vor (niedrige Messwerte), wobei einige Messstationen möglicherweise höhere Werte als die anderen aufweisen.

2 Umwandlung in funktionale Daten

Der Ausgangspunkt bei der funktionalen Datenanalyse ist die Annahme, dass die diskreten Daten „Momentaufnahmen" einer stetigen Funktion sind. Somit kann eine Zeitreihe, basierend auf den Luftschadstoffen, als Realisierung eines stetigen Prozesses betrachtet werden, die zu diskreten Zeitpunkten gemessen wird (vgl. Ignaccolo/Ghigo/Giovenali 2008, S. 675). Betrachtet man den Luftschadstoff L, so wird dieser an n_L Messstellen erhoben. Die diskreten Messwerte y_{ij}, $i = 1, \ldots, n_L$, $j = 1, \ldots, 1096$ können mit dem Modell[6]

$$y_{ij} = x_i(t_j) + \varepsilon_{ij} \qquad \varepsilon_{ij} \overset{iid}{\sim} F(0,\sigma^2), \text{ mit } i = 1, \ldots, n_L, j = 1, \ldots, 1096$$

beschrieben werden, mit (entsprechend glatten) Funktionen x_i und einem Zufallsfehlern ε_{ij}. Diese stetigen Funktionen x_i müssen in weiterer Folge geschätzt werden. Für nichtperiodische Daten eignen sich vor allem Polynom-Splines, wobei sich in der Praxis vor allem kubische B-Splines bewährt haben. Es stellt sich nun die Frage, wie viele Knoten in welchen Abständen für die Approximierung gewählt werden sollen. Ignoccolo, Ghigo und Giovenali (2008) wählen eine äquidistante Knotenmenge, und zwar jeweils einen pro Monat (vgl. Ignoccolo/Ghigo/Giovenali 2008, S. 676). Eine andere Möglichkeit besteht darin, die Knotenmenge fest zu wählen, und zwar jeweils zu Beginn eines Monats. Damit erhält man eine Knotenmenge, die nicht mehr gleichmäßig verteilt ist, sondern kleine Unterschiede aufweist. Abbildung 4.2.1 zeigt zwei vollständige Basen für den

[6] Fehlende Werte werden durch die jeweiligen Stationsmittelwerte ersetzt.

Erhebungszeitraum $T = (1, \ldots, 1096)$. In der oberen Darstellung sind die mittleren Basisfunktionen ident, während bei der unteren Graphik geringe Unterschiede in den Basisfunktionen erkennbar sind, und zwar in den Monaten Februar, da diese Intervalle am kleinsten sind.

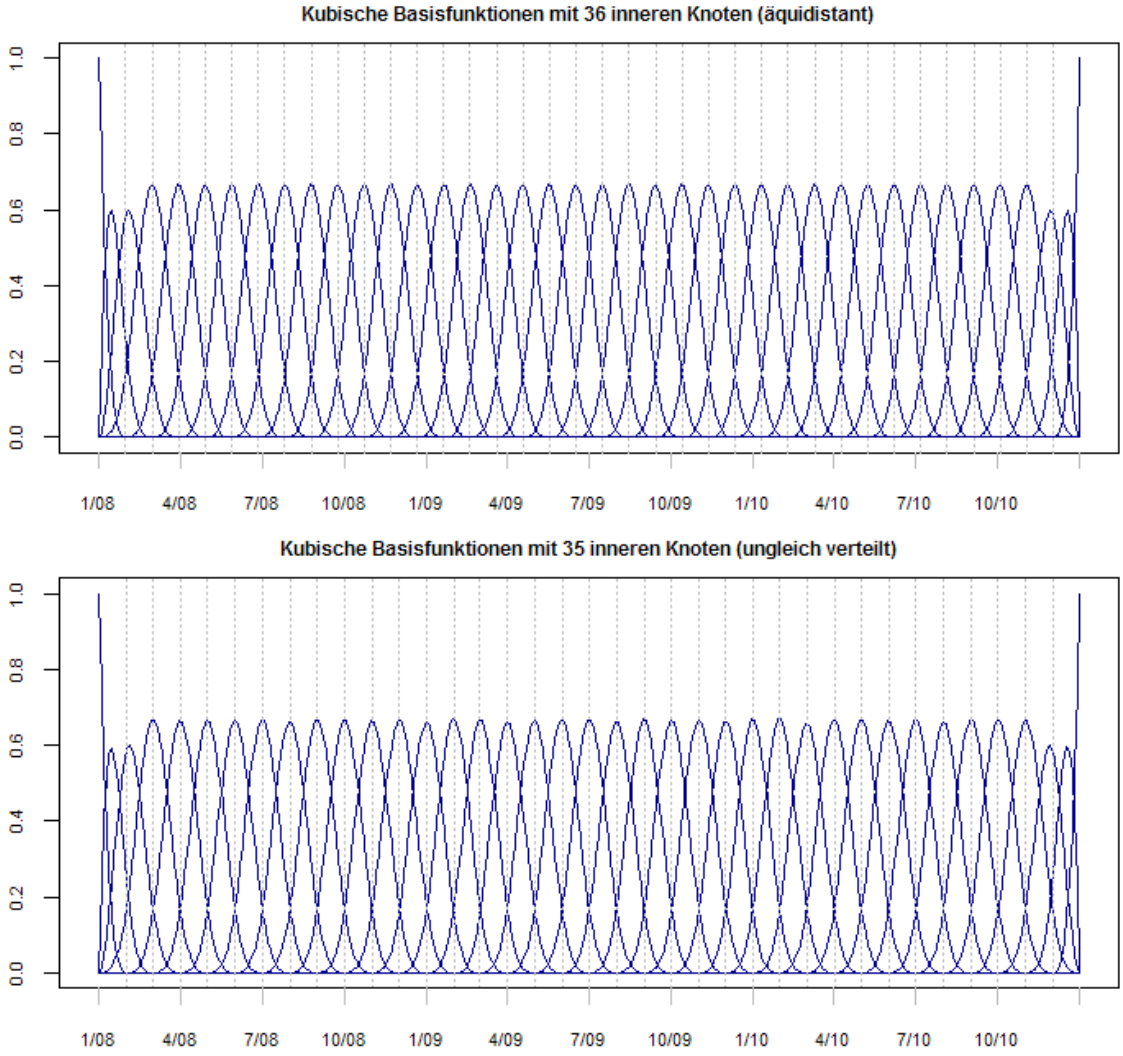


Abb. 4.2.1. Vollständige Basen für $T = (1, \ldots, 1096)$ mit äquidistanter bzw. ungleichmäßig verteilter Knotenmenge

Für die folgende Untersuchung ist es von Vorteil, wenn zur Kurvenapproximation derselbe Polynomgrad und dieselbe Knotenmenge für jeden einzelnen Luftschadstoff gewählt werden. Damit wird für die Schätzung jeweils dieselbe Basis verwendet. Das macht vor allem deswegen Sinn, da sich die Partitionierung auf die geschätzten Splinekoeffizienten bezieht.

Im Folgenden werden für die Luftschadstoffe die Kurven pro Messstation geschätzt. Für Feinstaub werden die Approximationen mit verschiedenen Basen, und zwar mit vierteljährlichen, monatlichen und vierzehntägigen Knoten, ausführlicher betrachtet. Für die

anderen Schadstoffe werden die Informationen zusammengefasst. Die Funktionen der Luftschadstoffe pro Messstation finden sich in Anhang B.1.

2.1 Luftschadstoff Feinstaub PM_{10} [µg/m³]

Die Funktionen x_i, $i = 1, \ldots, 25$, für die 25 Messstellen werden mithilfe von sogenannten Regression-Splines geschätzt. Penalisierungen, um eine bessere „Glattheit" zu erzeugen sind nicht notwendig. Diese werden vor allem dann eingesetzt, wenn die Anzahl der Basisfunktionen im Vergleich zur Anzahl der beobachteten Werte sehr groß ist (vgl. Ramsay/Hooker/Spencer 2009, S. 62). Bei dieser Untersuchung liegen die Anzahl der beobachteten Werte zwischen 413 und 1096, die Anzahl der Basisfunktionen bei 16, 22, 40 oder 76. Abbildungen 4.2.2 bis 4.2.4 zeigen Approximationen der Funktionen für den Luftschadstoff Feinstaub (PM_{10})[7].

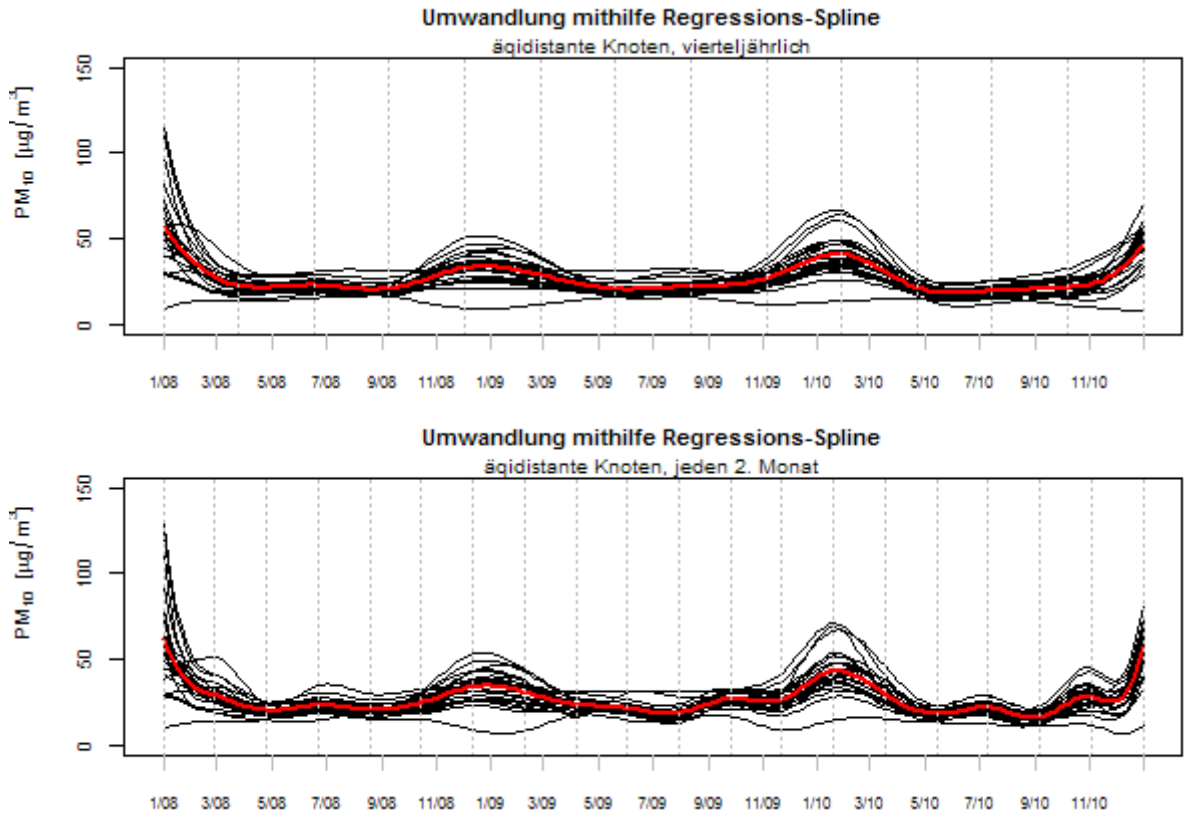


Abb. 4.2.2. PM_{10}: Approximation mithilfe von Regressionsspline bei 12 inneren Knoten (obere Graphik) bzw. 22 inneren Knoten (untere Graphik)

[7] Einige ausgewählte Routinen für die Umsetzung mithilfe der Statistik Software *R* befindet sich in Anhang C.

Bei einer Basisfunktion mit 12 inneren Knoten[8] bzw. 22 inneren Knoten[9] sind die Funktionen sehr glatt und nicht sehr ausgeprägt (siehe Abbildung 4.2.2). Insgesamt gibt es zwei klar erkennbare Gipfel, jeweils in der Wintersaison zwischen November und Februar. Jedoch weisen die Funktionen im Intervall *T* wenig spezifische Eigenschaften auf, daher könnte eine größere Anzahl an Knoten für eine genauere Schätzung interessant sein.

Bei einer Wahl von 72 inneren Knoten[10], also zwei pro Monat, sind die Funktionen x_i sehr detailliert, mit vielen Höhen und Tiefen (siehe Abbildung 4.2.3). Im Vergleich zu den vorherigen Approximationen (mit 16 bzw. 26 Basisfunktionen) wirken diese Funktionen „unruhig", d.h. die Genauigkeit geht zulasten der „Glattheit". Zusätzlich fällt eine lineare Funktion auf, und zwar entspricht diese den Messwerten von Graz Mitte ($i = 22$). Für diese Messstation gibt es keine Daten zwischen Februar 2008 und Jänner 2010. Diese fehlenden Werte werden durch den Stationsmittelwert ersetzt. Bei anderen Messstationen sind die Lücken kleiner und umfassen im Höchstfall zwei Monate.

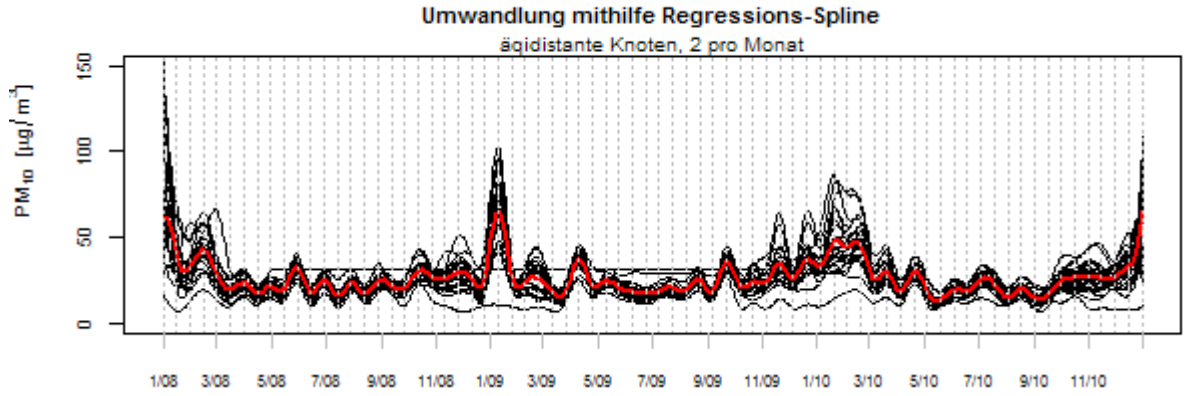


Abb. 4.2.3. PM_{10}: Approximation mithilfe von Regressionsspline bei 72 inneren Knoten

[8] Die 12 Knoten befinden sich an den Stellen {85.23, 169.46, 253.69, 337.92, 422.15, 506.39, 590.628, 674.85, 759.08, 843.31, 927.54, 1011.77}.
In den Abbildungen sind die Knoten (mit dem Start- bzw. Endknoten) jeweils durch horizontale punktierte Linien eingezeichnet. Für eine leichtere „Lesbarkeit" bzw. Interpretation ist die horizontale Achse in Monate eingeteilt.

[9] Die 22 Knoten befinden sich an den Stellen {58.63, 116.26, 173.90, 231.53, 289.16, 346.79, 404.42, 462.05, 519.68, 577.32, 634.95, 692.58, 750.21, 807.84, 865.47, 923.11, 980.74, 1038.37}

[10] Die 72 inneren Knoten befinden sich bei {16, 31, 46, 61, 76, 91, 106, 121, 136, 151, 166, 181, 196, 211, 226, 241, 256, 271, 286, 301, 316, 331, 346, 361, 376, 391, 406, 421, 436, 451, 466, 481, 496, 511, 526, 541, 556, 571, 586, 601, 616, 631, 646, 661, 676, 691, 706, 721, 736, 751, 766, 781, 796, 811, 826, 841, 856, 871, 886, 901, 916, 931, 946, 961, 976, 991, 1006, 1021, 1036, 1051, 1066, 1081}

Abbildung 4.2.4 zeigt ähnliche Kurven. Einige kleinere Unterschiede kann man in den Spitzen und in den Tälern ausmachen, die sich aufgrund der Position der Knoten[11] ergeben. Bei der äquidistanten Knotenmenge liegt jeweils ein Knoten inmitten des jeweiligen Monats, bei der ungleichmäßig verteilten Menge entsprechen die Knoten den Monatsersten.

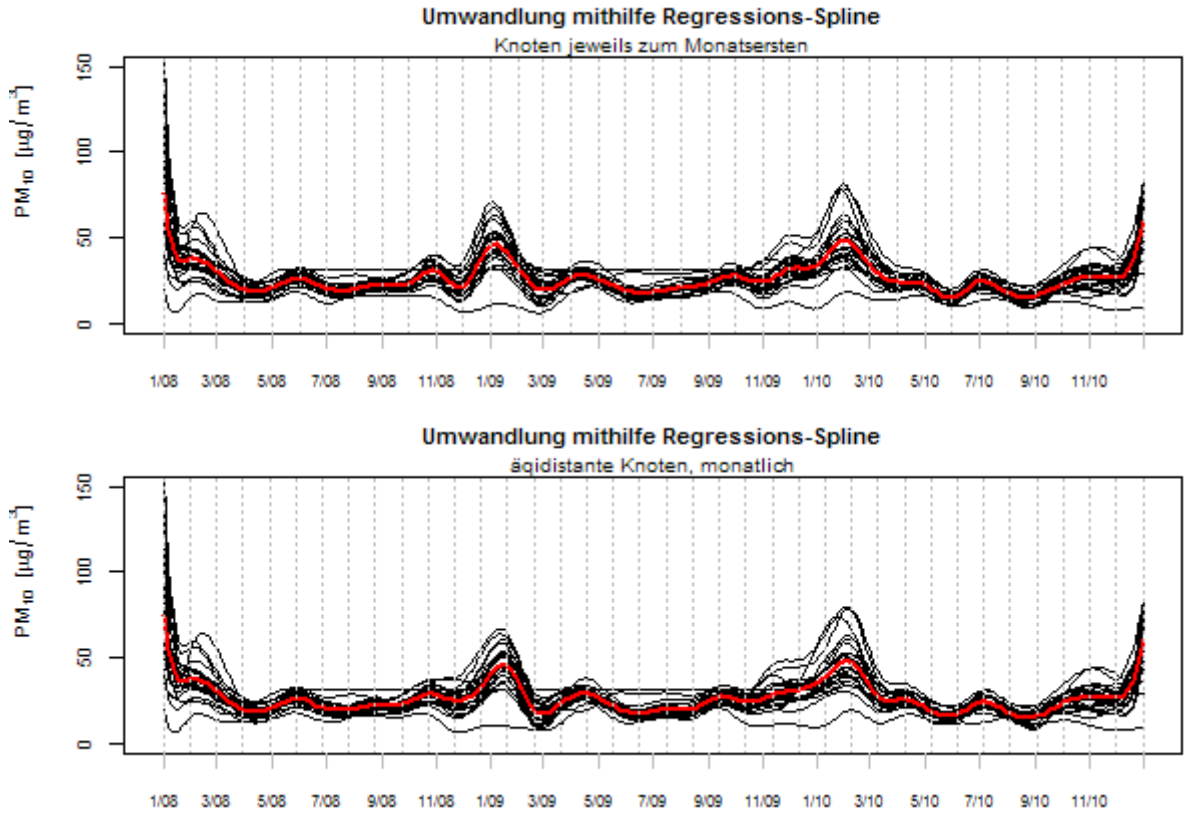


Abb. 4.2.4. PM_{10}: Approximation mithilfe von Regressionssplines bei 36 inneren Knoten (obere Graphik) und mit Knoten am Monatsersten (untere Graphik)

Beim Vergleich der verschiedenen Basen wird deutlich, dass die Anpassung bei einer Erhöhung der Knotenzahl „besser" wird, da die Funktionen mehr Informationen „verarbeiten". Dafür ist mit einem Verlust der „Glattheit" zu rechnen. Die Übergänge sind zwar aufgrund der Spline-Bedingungen gewährleistet, aber die Funktion als gesamtes wird „unübersichtlicher".

Um die Wahl einer Basis zu erleichtern, werden die Approximationen mit verschiedenen Basen für einen Messort gemeinsam mit den Messwerten y_j, $j = 1, \ldots, 1096$, betrachtet.

[11] äquidistante Knotenmenge: {30.59, 60.19, 89.78, 119.38, 148.97, 178.57, 208.16, 237.76, 267.35, 296.95, 326.54, 356.14, 385.73, 415.32, 444.92, 474.51, 504.11, 533.70, 563.30, 592.89, 622.49, 652.08, 681.68, 711.27, 740.86, 770.46, 800.05, 829.65, 859.24, 888.84, 918.43, 948.03, 977.62, 1007.22, 1036.81, 1066.41}
Knoten am Monatsersten: {1, 32, 61, 92, 122, 153, 183, 214, 245, 275, 306, 336, 367, 398, 426, 457, 487, 518, 548, 579, 610, 640, 671, 701, 732, 760, 791, 822, 852, 883, 913, 944, 975, 1005, 1036, 166, 1097}

Dafür eignet sich die Messstelle Graz Süd ($i = 24$), da diese Station keine fehlenden Werte hat und es insgesamt 1096 Beobachtungen gibt (siehe Abbildung 4.2.5).

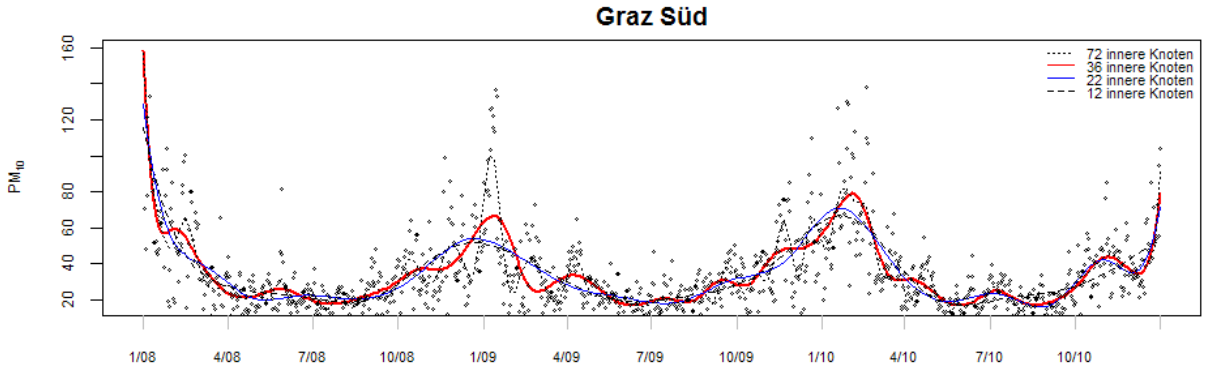


Abb. 4.2.5. PM_{10}: Approximationen der Funktion mit unterschiedlichen Basen für die Messstation Graz Süd

Bei den Messwerten gibt es insgesamt vier Peaks, und zwar am Anfang und am Ende der Messperiode und zwei dazwischen. Diese entsprechen hohen Beobachtungswerten in den Wintermonaten (Oktober bis April). Die inneren Maxima werden von allen drei Kurven wiedergegeben, sind jedoch unterschiedlich stark ausgeprägt. Die kleinsten Maxima zeigt, wie zu erwarten, die Approximation mit 12 inneren Knoten (blaue Kurve). Diese scheint auch andere kleinere Hoch- bzw. Tiefpunkte nicht anzugleichen. Dafür ist die Schätzung mithilfe von 72 inneren Knoten (schwarze Kurve) an einigen Stellen stark flukturierend. Sie weist deutlich mehr (lokale) Maxima und Minima als die beiden anderen Funktionen auf. Die rote Kurve stellt die Approximation mithilfe von 36 inneren Knoten dar. Diese Funktion enthält sowohl größere als auch kleinere Extrema und hat einen glatten Verlauf.

2.2 Luftschadstoff Ozon O_3 [µg/m³]

Der Luftschadstoff Ozon (O_3) wird nicht an allen 25 Messstellen erhoben, sondern an den 10 Stationen Deutschlandsberg, Fürstenfeld, Judenburg, Liezen, Masenberg, Mürzzuschlag, Voitsberg, Weiz, Graz Nord und Graz Süd Tiergartenweg. Somit müssen zehn Funktionen $x_i(t)$, $i = 1, \ldots, 10$, geschätzt werden. Bei Ozon tauchen die höchsten Werte in den Sommermonaten (Mai bis September) und die niedrigsten in den Wintermonaten (Oktober bis April) auf.

Wie beim Feinstaub ist die Schätzung mithilfe von 16 Basisfunktionen zu grob, um die Daten gut darzustellen (siehe Abbildung 4.2.6, erste Graphik). Die Kurven zeigen zwar die entsprechenden Maxima und Minima im Sommer bzw. im Winter, jedoch gibt es da-

zwischen kaum Variationen. Die Abstände der Knoten scheinen zu groß zu sein, um mehr Informationen aufzunehmen.

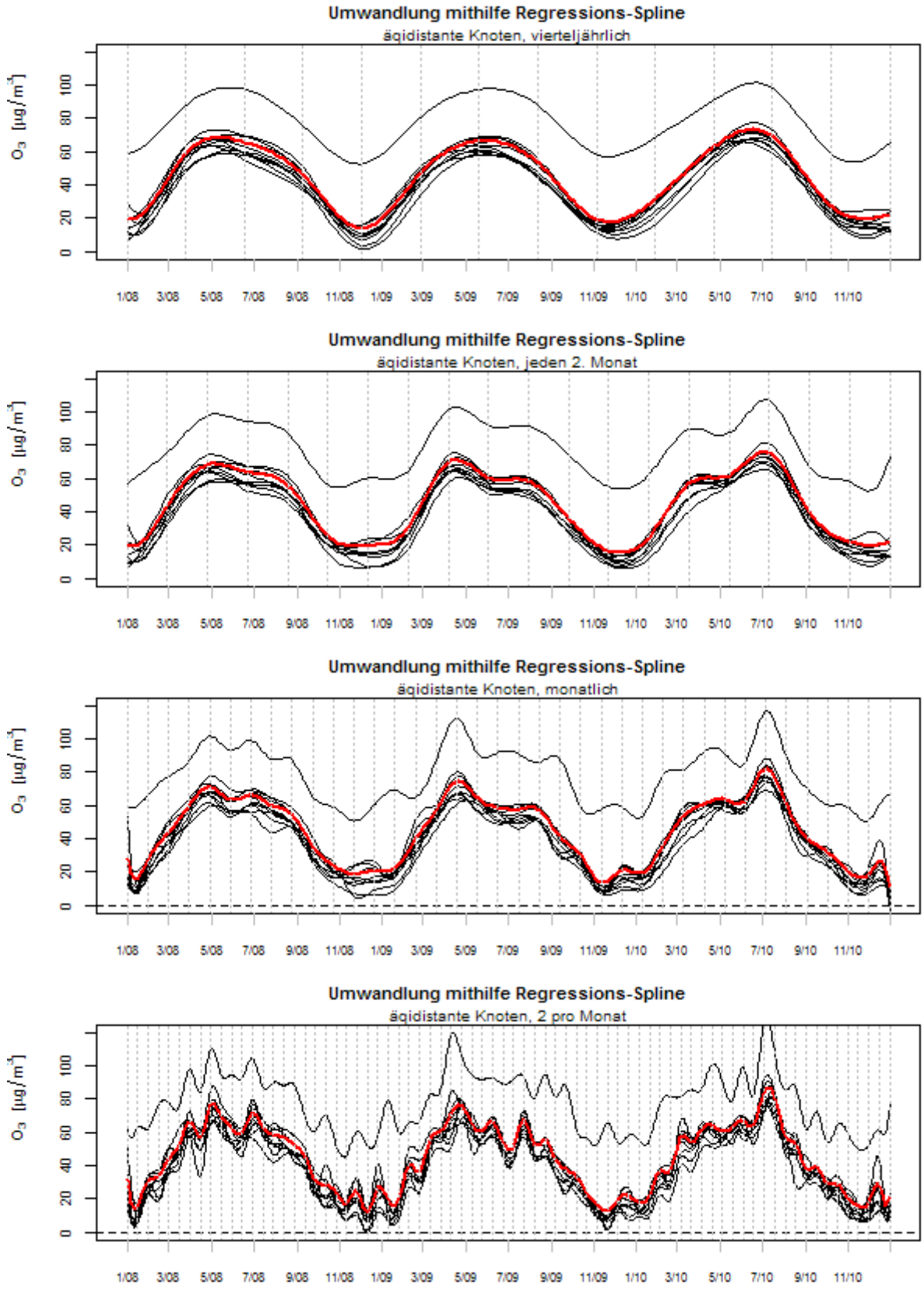


Abb. 4.2.6. O_3: Approximation mithilfe von Regressionsspline mit 12 (erste Graphik), 18 (zweite Graphik), 36 (dritte Graphik) und 72 (vierte Graphik) inneren Knoten

Eine Schätzung mithilfe von Knoten, die alle zwei Monate gesetzt werden, zeigt schon deutlichere Charakteristika der Funktionen (siehe Abbildung 4.2.6, zweite Graphik). Man erkennt bereits, dass die extremen Werte nicht über die gesamten Sommermonate auftreten, sondern dass es auch hier Spitzenwerte gibt (April/Mai 2009 und Juni/Juli 2010). Werden die Knoten enger gesetzt, und zwar alle vierzehn Tage (Basis zu 72 inneren Knoten), werden sehr viele Basisfunktionen verwendet und somit die Beobachtungswerte recht genau abgebildet (siehe Abbildung 4.2.6, vierte Graphik). Die Ausprägungen der Kurven bei dieser Approximation wirken nicht so unruhig wie bei Feinstaub, jedoch sind sie noch immer sehr flukturierend.

Der Unterschied bei einer Approximation mit 38 bzw. mit 37 Basisfunktionen ist auch bei den Daten des Luftschadstoffes Ozon gering, so dass die Schätzung mit einer äquidistanten Knotenwahl ausreichend ist. Diese Funktionen zeigen die Extrema in den Sommermonaten und Charakteristiken, die während der einzelnen Monate auftreten (siehe Abbildung 4.2.6, dritte Graphik).

2.3 Luftschadstoff Schwefeldioxid SO_2 [µg/m³]

Neunzehn der ausgewählten Messstationen haben über den Untersuchungszeitraum Daten zum Luftschadstoff Schwefeldioxid (SO_2) erhoben, und zwar Bruck, Deutschlandsberg, Fürstenfeld, Judendorf, Kapfenberg, Knittelfeld, Köflach, Leibnitz, Leoben/Göß, Leoben/Donawitz, Liezen, Masenberg, Niklasdorf, Peggau, Straßengel, Voitsberg, Graz Nord, Graz Süd und Graz West. Von den Messorten Kapfenberg und Leibnitz liegen nur vereinzelt Messwerte vor (70 Messwerte bzw. 122 Messwerte). Die Messstellen Leoben/Göß, Peggau und Graz West haben Beobachtungen von ca. 2 Jahren.

Im Folgenden werden nun diese neunzehn Funktionen $x_i(t)$, $i = 1, \ldots, 19$, mithilfe verschiedener Basen geschätzt. Das erste, das ins Auge fällt, ist die Kurve der Messstation Straßengel. Diese Funktion liegt über den anderen Kurven und weist zudem stärker ausgeprägte Extrema auf (siehe Abbildungen 4.2.7 und 4.2.9).

Betrachtet man die geschätzte Funktion gemeinsam mit den beobachteten Werte, erkennt man Unterschiede zu den der übrigen Messstationen[12].

[12] Die Funktionen mit den Beobachtungswerten befinden sich in Anhang B.1.

Der Mittelwert für Schwefeldioxid über alle Stationen hinweg beträgt 2.61 [µg/m³], wobei die meisten Stationsmittelwerte zwischen 1.55 und 4.67 [µg/m³] liegen. Somit ist der Mittelwert für die Messstation Straßengel mit 10.92 [µg/m³] deutlich höher.

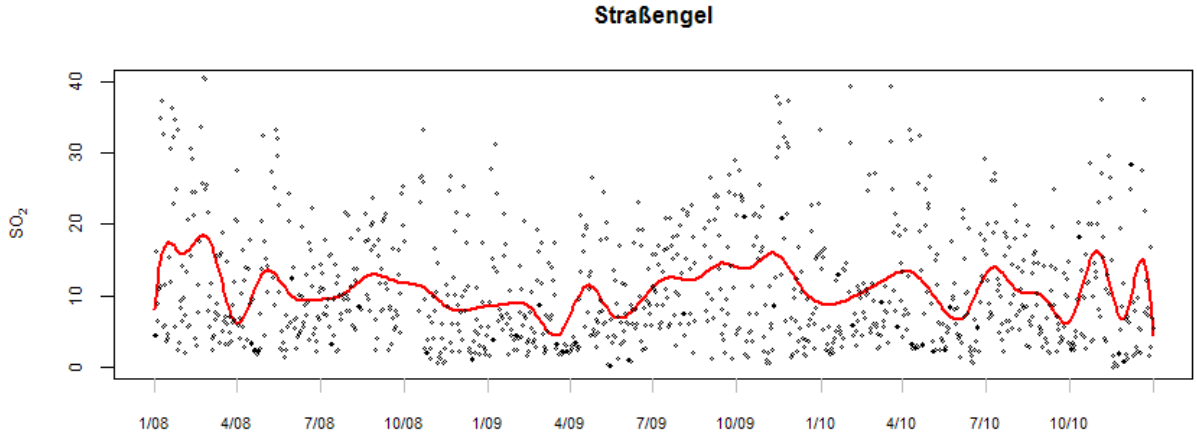


Abb. 4.2.7. SO_2: Die Funktion der Messstation Straßengel mit den beobachteten Werten

Mithilfe der Verteilung der Tagesmittelwerte wird veranschaulicht, wie stark die Messwerte bei Straßengel von denen der anderen Stationen abweichen (siehe Abbildung 4.2.8). Der interquartile Bereich ist bei Straßengel deutlich größer und auch der maximale Wert (65.58 [µg/m³]) liegt weit über den anderen. Bis auf Judendorf (51.44 [µg/m³]) liegen die maximalen Werte unter 23 [µg/m³], zum Teil sogar unter 10 [µg/m³].

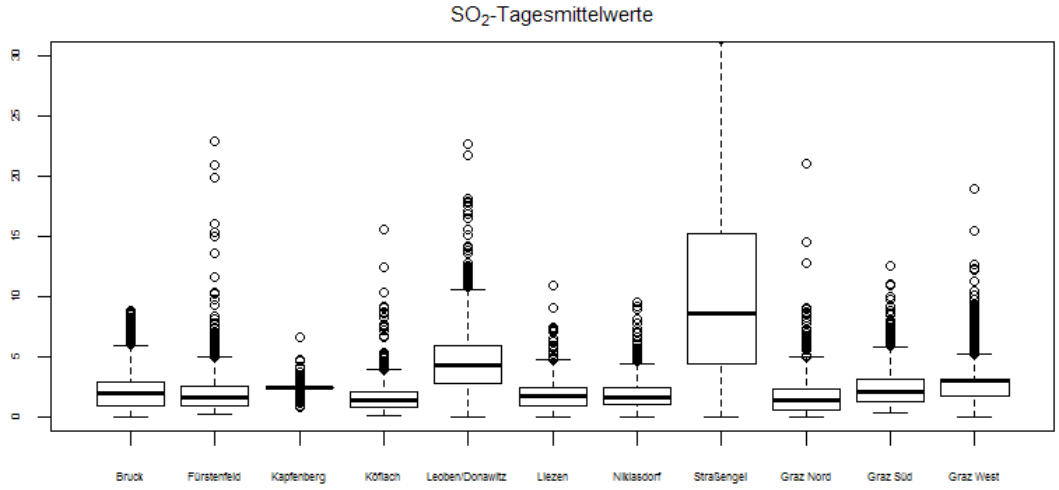


Abb. 4.2.8. Verteilung der SO_2-Tagesmittelwerte für 11 Messstationen

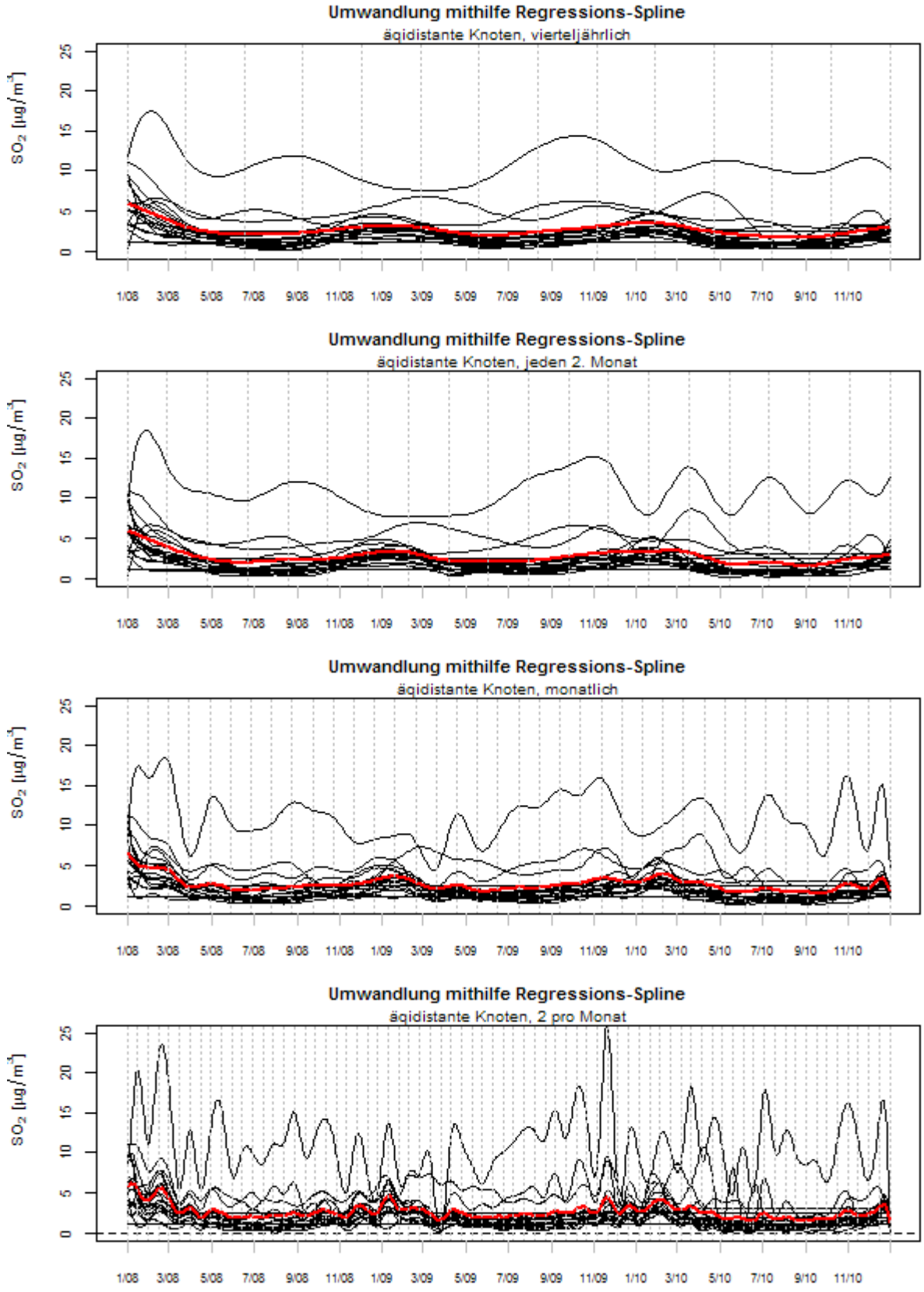


Abb. 4.2.9. SO_2: Approximation mithilfe von Regressionsspline mit 12 (obere Graphik), 36 (mittlerer Graphik) und 72 (untere Graphik) inneren Knoten

2.4 Luftschadstoff Stickstoffdioxid NO_2 [µg/m³]

Wie bei Feinstaub haben alle 25 Messstellen Beobachtungen für den Luftschadstoff Stickstoffdioxid registriert. Zusätzlich sind bei jeder Station, außer Graz Mitte, ausreichend Daten vorhanden (mehr als 1000 Beobachtungen).

Betrachtet man die Approximationen der Funktionen $x_i(t)$, $i = 1, \ldots, 25$, mit den verschiedenen Basen, ergibt sich eine ähnliche Situation wie bei den anderen Luftschadstoffen (siehe Abbildung 4.2.11). Die Schätzung mit einer geringeren Anzahl an Knoten ist zwar sehr glatt, aber diese Funktionen haben wenige Ausprägungen. Deutlicher treten die Unterschiede im Jahresverlauf mit einer höheren Anzahl an Knoten hervor.

Bei allen Approximationen fällt eine Kurve auf, und zwar jene nahe der horizontalen Achse, die beinahe den Verlauf einer konstanten Funktion hat. Betrachtet man die Funktion der Messstation Masenberg gemeinsam mit deren Beobachtungswerten, sind Maxima und Minima erkennbar (siehe Abbildung 4.2.10). Daraus wird ersichtlich, dass die Messwerte dieser Messstation deutlich unter denen der anderen Stationen liegen. So ist der maximale Wert (24.9 [µg/m³]) bzw. das arithmetische Mittel (4.0 [µg/m³]) dieses Messortes wesentlich geringer als der Gesamtmaximalwert (137.5 [µg/m³]) bzw. der Gesamtmittelwert (21.8 [µg/m³]) über alle Stationen hinweg.

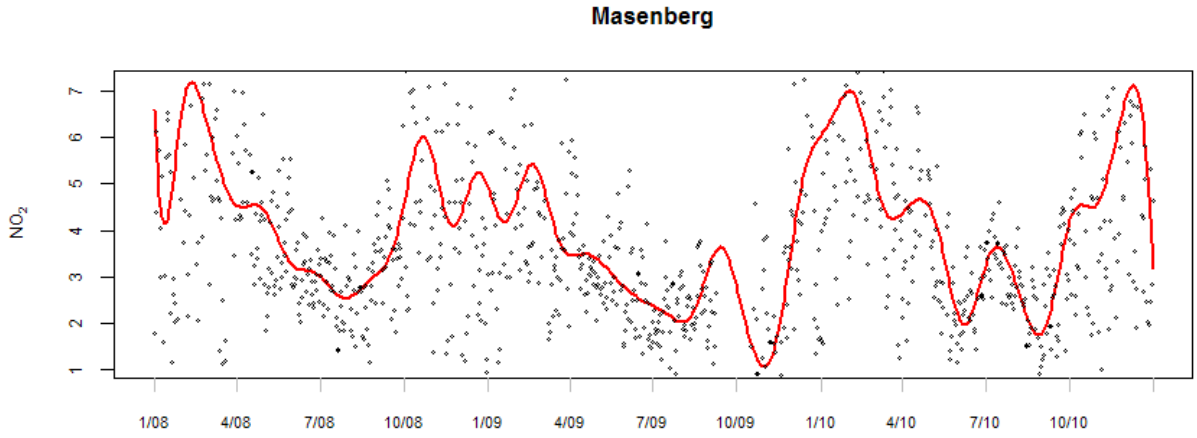


Abb. 4.2.10. NO_2: Die Funktion der Messstation Masenberg mit den beobachteten Werten

Eine andere Kurve, und zwar jene von Graz Don Bosco, ist aufgrund der großen Werte exponiert. Die Funktionswerte sind mindestens 15 [µg/m³] größer als die der übrigen Funktionen. Bei dieser Station findet man auch den Maximalwert. Das arithmetische Mittel (49.65 [µg/m³]) liegt zudem deutlich über dem Gesamtmittelwert.

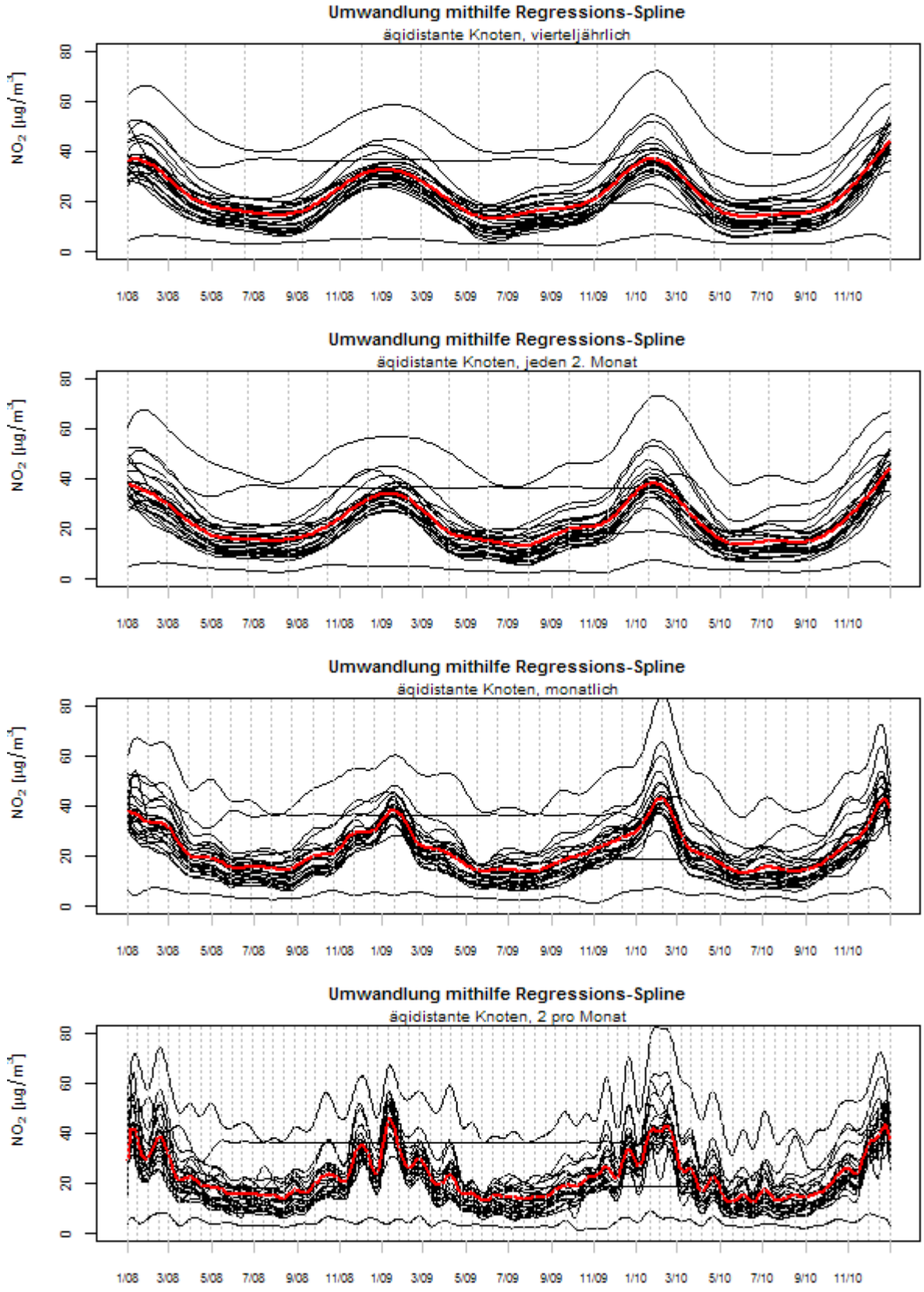


Abb. 4.2.11. NO_2: Approximation mithilfe von Regressionsspline mit 12 (erste Graphik), 18 (zweite Graphik), 36 (dritte Graphik) und 72 (vierte Graphik) inneren Knoten

3 Funktionales Clustern der Daten

Beim Clustern von funktionalen Daten wird der Verlauf der Funktion als Ausgangspunkt genommen, somit geht man von (geschätzten) Parametern aus, z. B. von den geschätzten Splinekoeffizienten oder den Knoten. Diese Parameter werden dann mit einem bekannten Verfahren, z. B. mit *k-means*-Verfahren oder *k-medoids*-Verfahren, geclustert. Beim *k-means*-Verfahren erhält man daraufhin die entsprechenden Koeffizienten der Funktionen, die die Zentren der Cluster darstellen. Beim *k-medoids*-Verfahren werden aus allen Parametern Repräsentanten für die Gruppierungen gewählt. Eine andere Möglichkeit der Partition besteht darin, die Rohdaten als Grundlage fürs Clustern zu verwenden und die gewählten Daten mithilfe der Basisfunktionen zu glätten.

Im Folgenden werden das *k-means*- und *k-medoids*-Verfahren für die vorliegenden Daten verwendet, wobei als Parameter die geschätzten Splinekoeffizienten gewählt werden. Als Vergleich wird das *k-means*-Verfahren auch auf die Rohdaten angewandt, um daraufhin die Funktionen für die Clusterzentren mithilfe der B-Spline-Basisfunktionen zu bilden.

3.1 Ergebnisse zu Feinstaub

3.1.1 Ergebnis aus *k-means* der geschätzten Splinekoeffizienten

Das *k-means*-Verfahren zählt zu den nichthierarchischen Clustermethoden, daher ist die Klassenanzahl vorher festzulegen. Eine Entscheidungshilfe stellt der sogenannte Screeplot dar, dabei wird eine variierende Anzahl an Klassen in Hinblick auf das zu minimierende Zielkriterium (*sum of squares within*) dargestellt. Ein Balkendiagramm eignet sich gut zur Veranschaulichung, weil das Zielkriterium mit der steigenden Anzahl der Klassen monoton fallend ist (vgl. Öllinger 2010, S. 36).

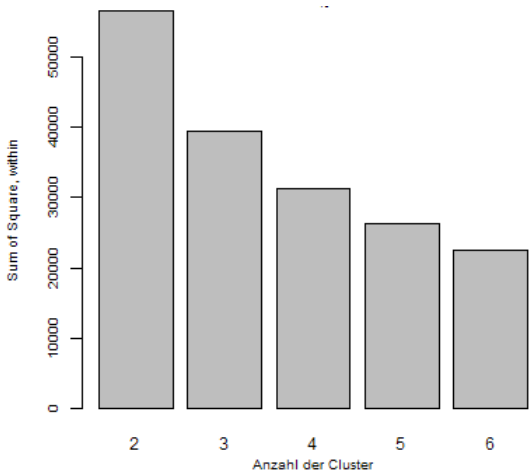


Abb. 4.3.1. PM_{10}: Screeplot für eine Clusteranzahl von $k = 2, \ldots, 6$

Als optimale Klassenanzahl wird jene gewählt, die zur vorherigen Anzahl die größtmögliche Differenz aufweist. Für den Luftschadstoff Feinstaub ergibt sich als optimale Klassenanzahl $k = 3$ (siehe Abbildung 4.3.1). Beim Clustern erhält man dann eine große

Gruppe mit insgesamt 16 Messstellen und zwei kleine Cluster mit jeweils 4 Stationen (siehe Abbildung 4.3.2), wobei die Differenzierung nach dem Niveau des Feinstaubs erfolgt.

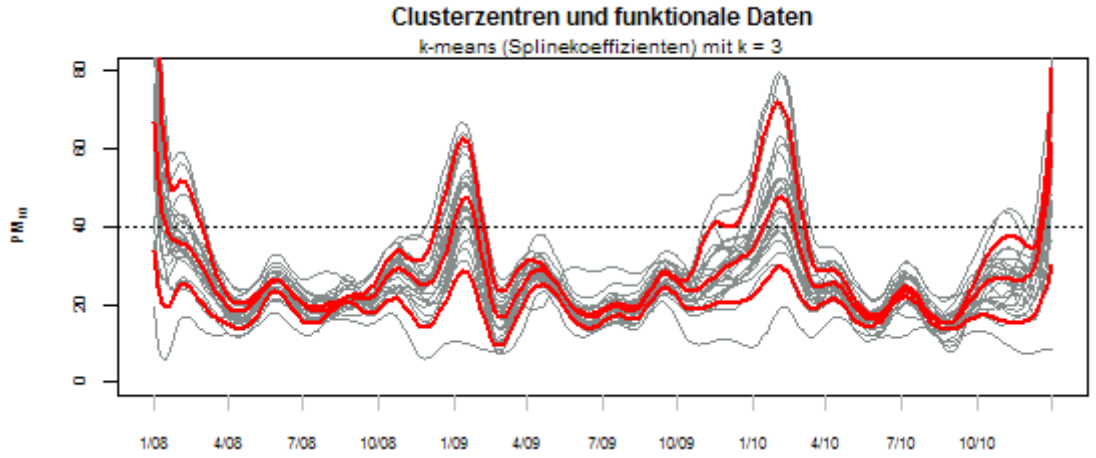


Abb. 4.3.2. PM_{10} – Clusterzentren (rote Funktionen) mit den funktionalen Daten der Messstationen bei einer Clusterung der geschätzten Splinekoeffizienten (*k-means*, *k = 3*).

Das dritte Cluster fasst jene Messstationen zusammen, die eher niedrige Feinstaubwerte über die drei Jahre gemessen haben (siehe Abbildung 4.3.3, untere Graphik). Auch sind die Maxima weniger stark ausgeprägt, besonders jenes zu Beginn des Jahres 2010. Zu diesen zählen die Messorte Judenburg, Liezen, Masenberg und Mürzzuschlag. Die Funktionen selbst weisen stärkere Unterschiede zueinander auf, als es bei den anderen Clustern der Fall ist.

Die Kurven der Stationen Leibnitz, Graz Don Bosco, Graz Süd und Graz Nord im ersten Cluster liegen über den anderen, d.h. hier sind die Messwerte höher mit ausgeprägten Spitzen (siehe Abbildung 4.3.3, obere Graphik). Es ist nicht überraschend, dass drei der vier Grazer Messstellen zu der Gruppe der hohen Feinstaubbelastung zählen.

Die Werte der übrigen 16 Stationen sind recht ähnlich, sie werden daher in einem Cluster zusammengefasst (siehe Abbildung 4.3.3, mittlere Graphik). Man erkennt deutlich, dass diese Funktionen kaum Unterschiede im Verlauf aufweisen.

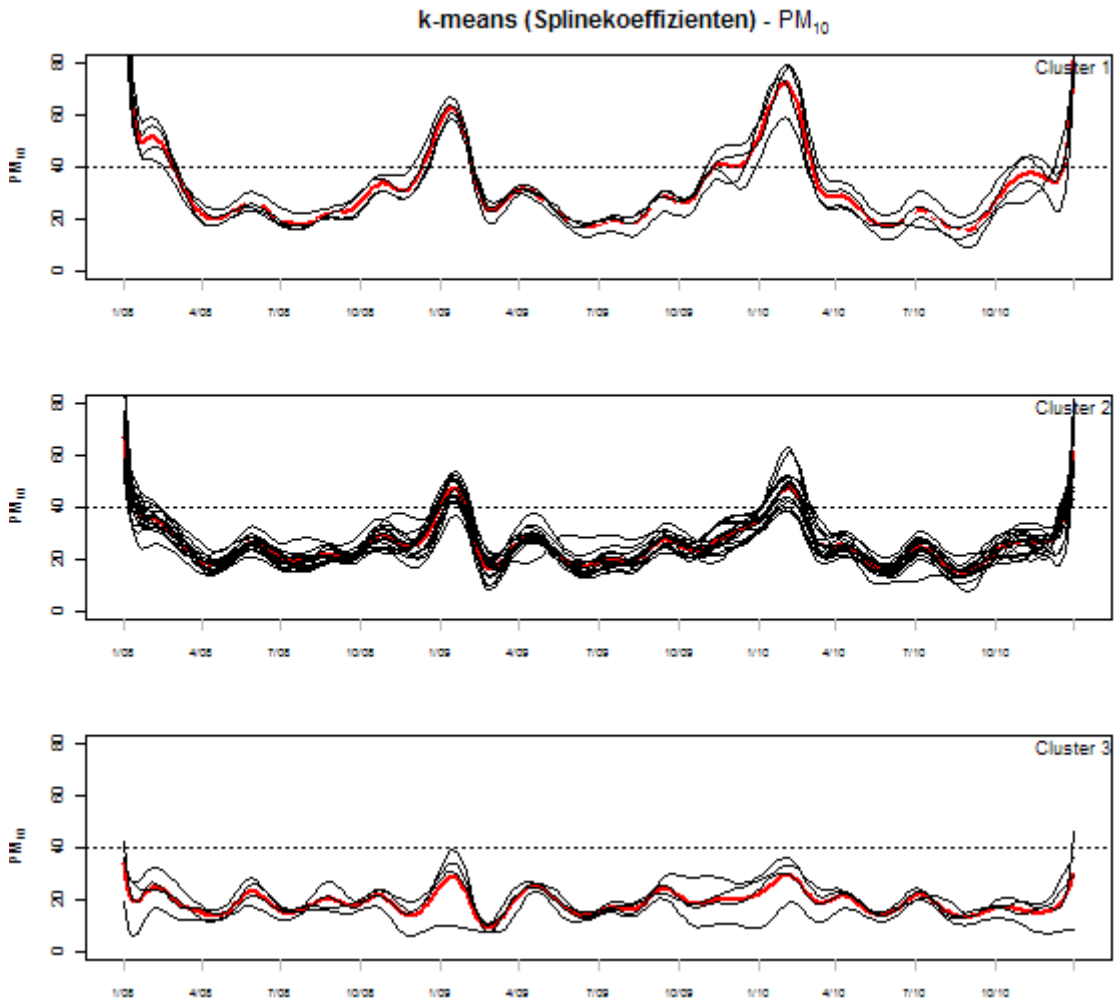


Abb. 4.3.3. PM_{10} – Cluster beim *k-means* der geschätzten Splinekoeffizienten (rote Funktion: Clusterzentrum)
Cluster 1: Leibnitz, Graz Don Bosco, Graz Süd, Graz West
Cluster 2: Bruck, Deutschlandsberg, Fürstenfeld, Judendorf, Kapfenberg, Knittelfeld, Köflach, Leoben/Donawitz, Leoben/Göß, Niklasdorf, Peggau, Straßengel, Voitsberg, Weiz, Zeltweg, Graz Nord
Cluster 3: Judenburg, Liezen, Masenberg, Mürzzuschlag

Bei einer Aufteilung in vier Gruppen bleiben die Cluster, in denen sich die Messstationen mit den höchsten bzw. den niedrigsten Messwerten befinden, gleich[13]. Das Cluster mit den mittleren Werten wird in eine Gruppe mit 11 und in eine mit 5 Messstationen aufgespalten. Dem kleineren Cluster werden Bruck, Kapfenberg, Leoben/Donawitz, Leoben/Göß und Niklasdorf zugeordnet. Diese Trennung ist jedoch nicht sehr scharf, da die durchschnittlichen Silhouette-Breiten[14] dieser beiden Cluster um 0.15 liegen.

Für das Clustern wurde Graz Mitte Gries aus dem Datensatz entfernt, da diese Messstation nur Daten von ungefähr einem Jahr erhoben hatte. Wird sie bei der Clusterung

[13] Tabellen mit der Zuordnung der Messstationen zu den Clustern bei unterschiedlicher Klassenanzahl befinden sich in Anhang A.2.
[14] Die Silhouette-Plots sind in Anhang B.3 zu finden.

berücksichtigt, ergibt sich kein Unterschied bei einer Clusteranzahl von *k* = 3. Erst ab einer Aufteilung in vier Gruppen kann man einen Unterschied wahrnehmen: Die Aufspaltung der großen Gruppe mit den „mittleren Messwerten" ergibt ein Cluster mit 15 Messstellen und ein Cluster mit der Station Graz Mitte Gries.

3.1.2 Ergebnis *aus k-means* der Rohdaten

Bei der Clusterung der Rohdaten durch das *k-means*-Verfahren und der Ermittlung der Clusterzentren mithilfe der B-Spline-Basisfunktionen ergibt sich mithilfe des Screeplot wiederum eine ideale Clusteranzahl von *k* = 3[15]. Die Aufteilung der Messstationen mit ihren funktionalen Daten ist jedoch eine andere als beim Clustern der geschätzten Splinekoeffizienten. Die Gruppen beinhalten dabei 12, 8 und 4 Messstationen (siehe Abbildung 4.3.4).

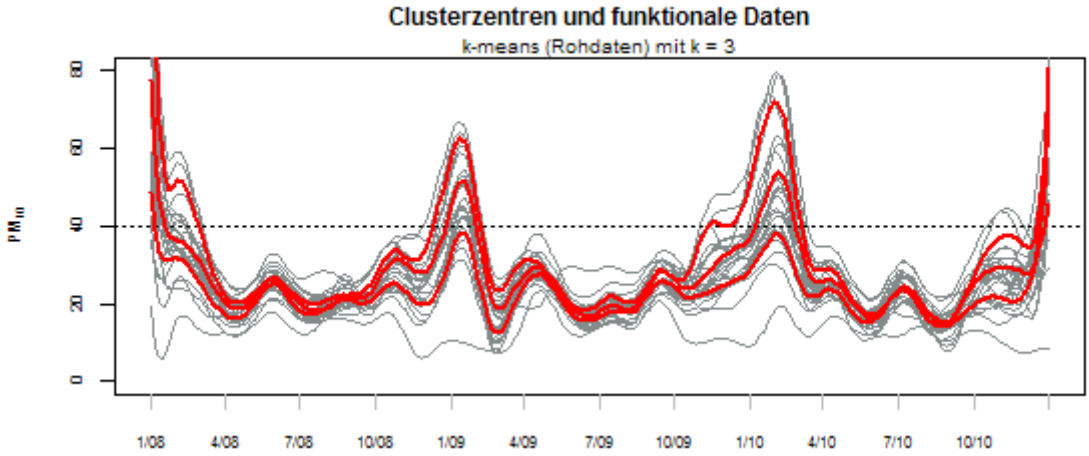


Abb. 4.3.4. PM_{10} – Clusterzentren (rote Funktionen) mit den funktionalen Daten der Messstationen bei einer Clusterung der Rohdaten (*k-means*, *k* = 3)

Die Größe des Clusters mit den Messstationen, die die höchsten Werte misst – in dem Fall Cluster 3 – enthält vier Objekte, und zwar Leibnitz, Graz Don Bosco, Graz Süd, Graz West (siehe Abbildung 4.3.5). Diese stimmt mit dem Cluster 1 der vorigen Berechnung überein. Jedoch unterscheiden sich die Zuordnungen für die beiden anderen Cluster. In der Gruppe mit den niedrigsten Messwerten sind nun mehr Messstationen enthalten (insgesamt 12), wobei die Anzahl der Objekte im Cluster mit den mittleren Werten halbiert wird. Die durchschnittlichen Silhouette-Breiten[16] der Cluster liegen bei 0.32, 0.31 und 0.15, um einiges geringer als bei der vorherigen Methode. Auch bei einer anderen Wahl der Klassenanzahl (*k* = 2, 4 oder 5) ist die durchschnittliche Silhouette-Breite jeweils

[15] Screeplots sind in Anhang B.2 zu finden.

geringer als jene bei einer Clusterung mit dem *k-means*-Verfahren der Splinekoeffizienten.

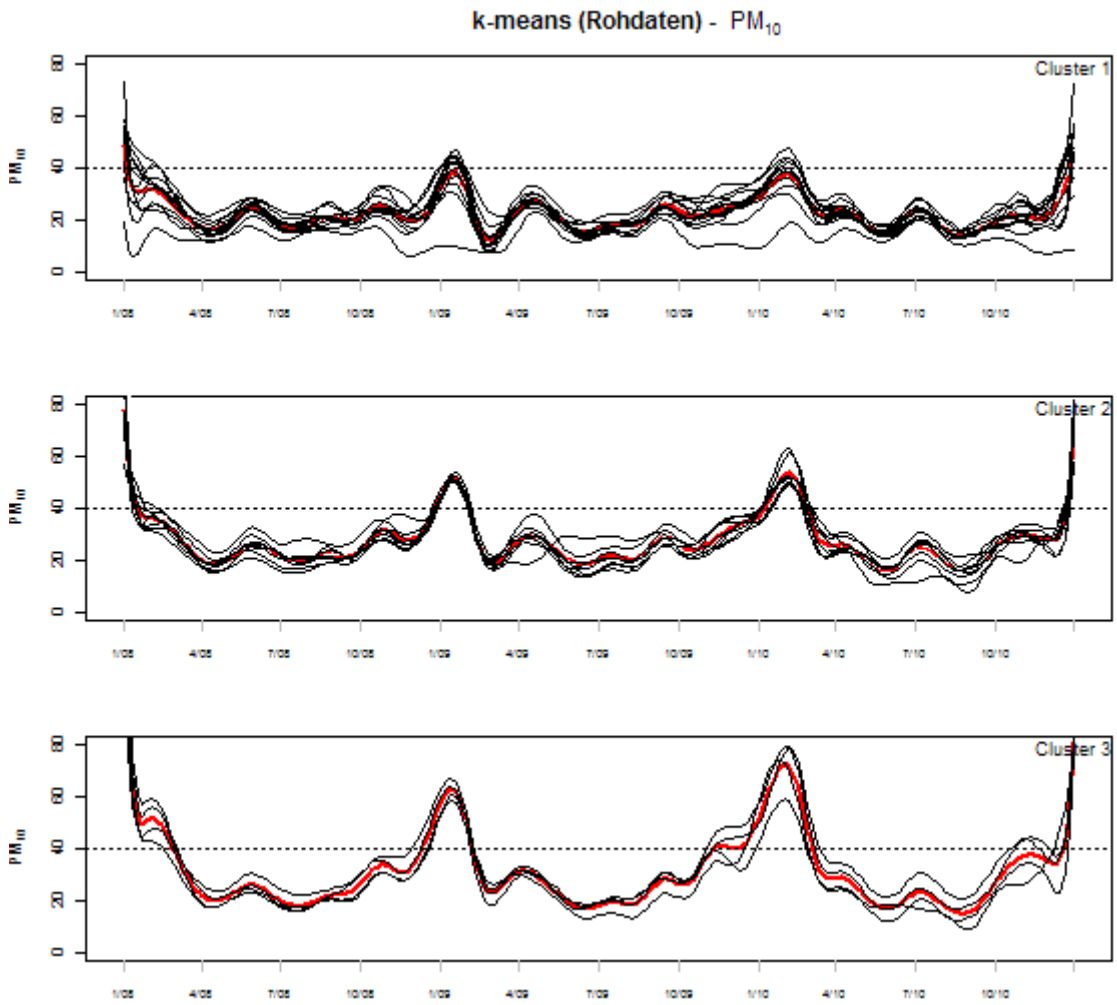


Abb. 4.3.5. PM_{10} – Cluster beim *k-means* der Rohdaten (rote Funktion: Clusterzentrum)
Cluster 1: Bruck, Judenburg, Kapfenberg, Knittelfeld, Leoben/Donawitz, Leoben/Göß, Liezen, Masenberg, Mürzuschlag, Niklasdorf, Straßengel, Zeltweg
Cluster 2: Deutschlandsberg, Fürstenfeld, Judendorf, Köflach, Peggau, Voitsberg, Weiz, Graz Nord
Cluster 3: Leibnitz, Graz Don Bosco, Graz Süd, Graz West

3.1.3 PAM aus geschätzten Splinekoeffizienten

Das *Partitioning Around Medoids*-Verfahren, kurz *PAM*, eignet sich besonders gut bei Daten, bei denen die Cluster als kugelförmig betrachtet werden. Ignoccolo, Ghigo und Giovenali (2008, S. 676ff.) gehen davon aus, dass Messstationen aus einer Region nahe beieinander liegen und somit auch Ähnlichkeiten bei den Messungen aufweisen. Die Messorte in der Steiermark sind zwar nicht so nahe beieinander, aber man kann Unter-

[16] Die Silhouette-Plots befinden sich in Anhang B.3.

schiede bei Messungen in der Obersteiermark und der südlichen Region erkennen. Daher kann möglicherweise das *PAM*-Verfahren interessante Informationen liefern

Beim *Partitioning Around Medoids* werden *k* Repräsentanten (Medoids) gewählt, um die herum die restlichen Objekte der Cluster zu finden sind. Diese Medoids sollen somit die „charakteristischen" Objekte eines Clusters darstellen. Dabei beginnt der Algorithmus mit einer (zufälligen) Auswahl an Repräsentanten und ersetzt diese in jedem Schritt durch „bessere". Aber auch hier muss im Vorfeld überlegt werden, wie viele Cluster es geben soll bzw. wie viele Repräsentanten gewählt werden. *PAM* selbst stellt dazu ein eigenes Instrument zur Verfügung, und zwar die sogenannte „silhouettes". Für jedes Objekt wird eine Kennzahl zurückgegeben, die angibt, wie gut es in dieses Cluster passt bzw. ob es eher zwischen zwei Cluster liegt. Damit hat man auch eine Vergleichsmöglichkeit für mehrere Partitionen (vgl. Rousseeuw 1987, S. 55). Um sich für eine bestimmte Anzahl von Repräsentanten zu entscheiden, werden in einem ersten Schritt die durchschnittlichen Silhouette-Breiten für verschiedene Klassenanzahl *k* betrachtet (siehe Abbildung 4.3.6). Für den Luftschadstoff Feinstaub ist die durchschnittliche Silhouette-Breite für eine Partitition in zwei Klassen deutlich höher als bei den anderen Aufteilungen (0.42).

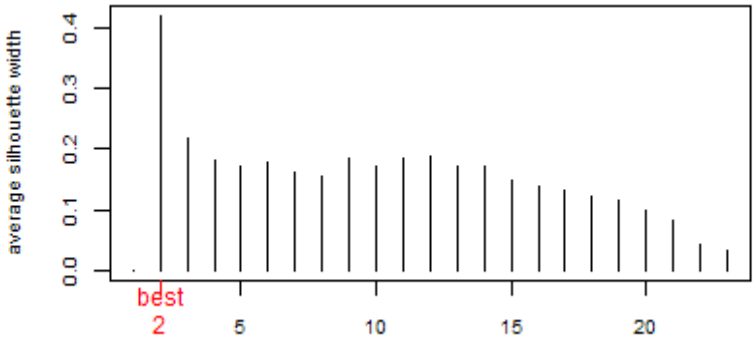


Abb. 4.3.6. PM10: durchschnittliche Silhouette-Breiten für schiedene Clusteranzahl k

Eine Gruppierung mit drei Repräsentanten halbiert bereits die durchschnittliche Silhouette-Breite (0.21). Bei einer Clusterung von 2 Klassen sind die durchschnittlichen Silhouette-Breiten pro Gruppe ebenfalls recht hoch (Cluster 1: 0.41, Cluster 2: 0.46). Außerdem sind die Silhouette-Breiten[17] für jede Messstation positiv, wobei Graz Nord und Deutschlandsberg jedoch sehr kleine Werte aufweisen (unter 0.2). Bei einer Aufteilung in drei Cluster gibt es bereits fünf Objekte, deren Zuordnung zu einer Gruppe nicht mehr eindeutig ist, und zwar Zeltweg, Leoben/Donawitz, Knittelfeld, Köflach und Straßengel. Zwei Stationen haben sogar einen negativen Silhouette-Wert (um –0.2), und zwar Fürstenfeld und Graz West. Daraus kann man schließen, dass diese sieben Messstationen zwischen zwei Cluster oder am Rand eines Clusters liegen.

[17] Tabellen mit den Silhouette-Breiten pro Messstation gibt es in Anhang A.3 und die entsprechenden Silhouette-Plots für die verschiedenen Clusterings in Anhang B.3.

Betrachtet man die Zuordnung der Messstationen bei dieser Clusterung, erkennt man, dass diese Partition eine Aufteilung zwischen dem Grazer Becken (Grazer Stationen, außer Graz Nord, und Leibnitz) und dem Rest der Steiermark vornimmt. Um eine bessere Vergleichsmöglichkeit zu den anderen Verfahren zu erhalten, wird auch für diese Methode eine Klassenanzahl von *k* = 3 gewählt (siehe Abbildung 4.3.7), auch wenn sich die durchschnittliche Silhouette-Breite deutlich von derjenigen mit zwei Gruppen unterscheidet.

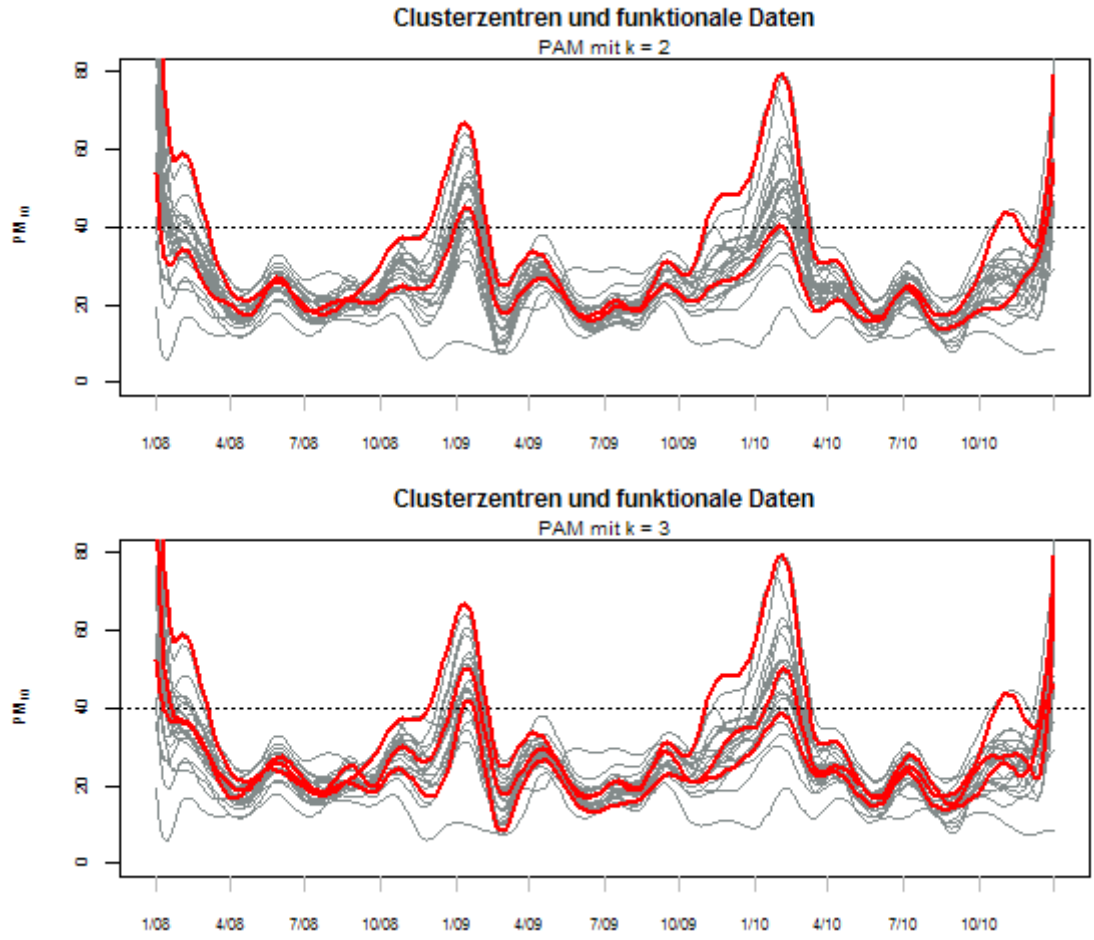


Abb. 4.3.7. PM_{10} – Clusterzentren (rote Funktionen) mit den funktionalen Daten der Messstationen bei *PAM* der geschätzten Splinekoeffizienten, für *k* = 2 (obere Graphik) und *k* = 3 (untere Graphik)

Das *PAM*-Verfahren wählt die Messstellen **Niklasdorf**, **Graz Nord** und **Graz Süd** als Repräsentanten der Cluster (siehe Abbildung 4.3.8). Graz Süd ist dabei eine „typische" Messstation für hohe Werte, der Verlauf ihrer Funktion unterscheidet sich nur geringfügig von jener der Station Graz Don Bosco. Die Silhouette-Breiten sind für die Grazer Stationen sehr hoch (0.52 bzw. 0.47) im Vergleich zu Leibnitz (0.06). Dieser Zusammenhang zeigt sich bereits in den Messdaten. Die Beobachtung an der Messstation Leibnitz sind ein wenig niedriger als die der beiden Grazer Stationen.

Bei einer Clusterung mit zwei Klassen befinden sich im Cluster mit den höchsten Werten Graz Nord, Graz Süd, Leibnitz und zusätzlich Graz West. Die restlichen Messstationen werden dem zweiten Cluster zugeordnet. Daher unterscheiden sich die ersten beiden Cluster bei einer Aufteilung auf drei Gruppen nicht sehr stark voneinander, was auch an den Silhouette-Breiten erkennbar ist.

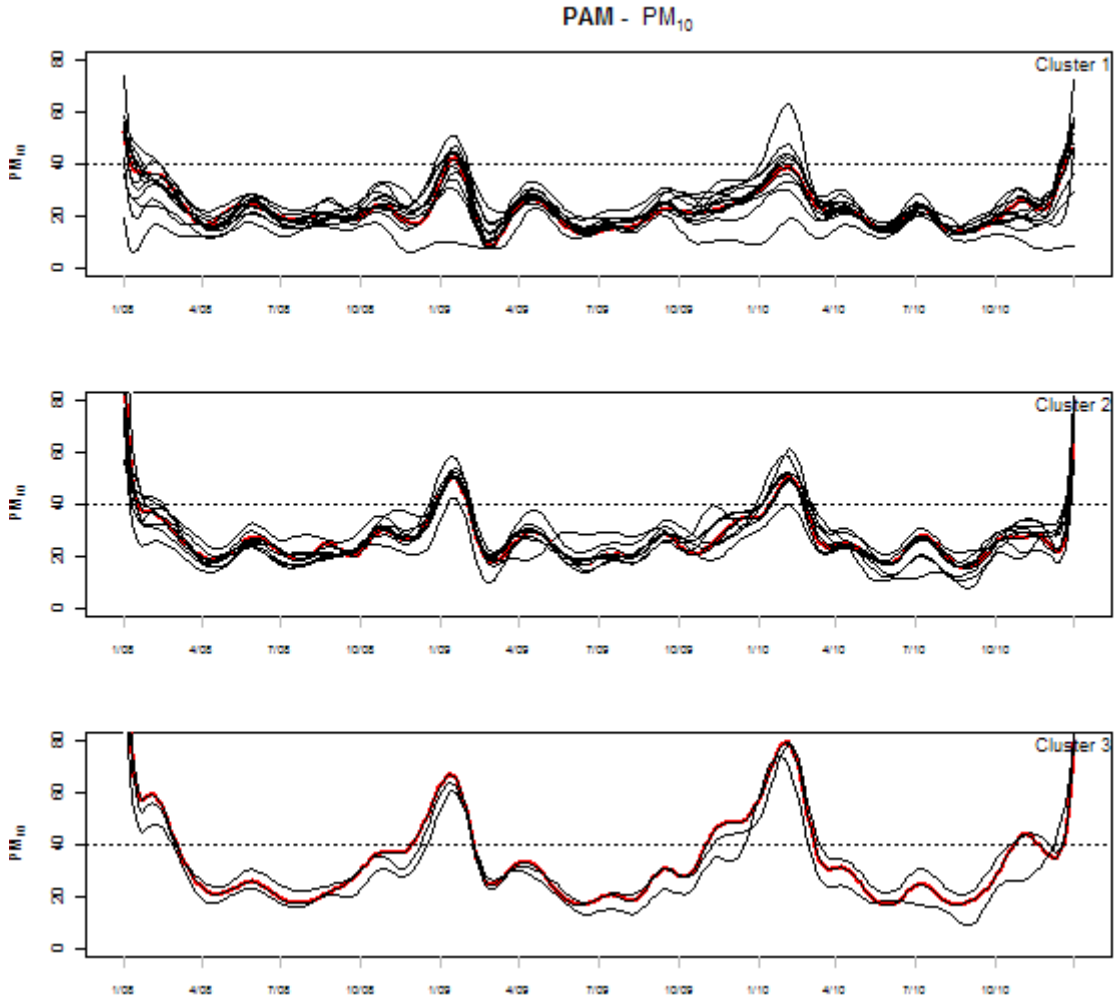


Abb. 4.3.8. PM_{10} – Cluster beim *k-means* der Rohdaten (rote Funktion: Clusterzentrum)
Cluster 1: Bruck, Fürstenfeld, Judenburg, Kapfenberg, Knittelfeld, Leoben/Donawitz, Leoben/Göß, Liezen, Masenberg, Mürzzuschlag, **Niklasdorf**, Zeltweg
Cluster 2: Deutschlandsberg, Judendorf, Köflach, Peggau, Straßengel, Voitsberg, Weiz, **Graz Nord**, Graz West
Cluster 3: Leibnitz, Graz Don Bosco, **Graz Süd**

3.1.4 Vergleich der Clusteranalyseverfahren

Interessant ist natürlich ein Vergleich der drei Methoden in Bezug auf die Zuordnung der Messstationen zu den einzelnen Clustern. Eine Möglichkeit zum Aufdecken von Gemeinsamkeiten bzw. Unterschieden bietet der *Randindex*[18], dabei werden zwei Partitionen in Bezug auf Übereinstimmung bei Paaren von Objekten betrachtet. Als Übereinstimmung gilt, wenn die beiden Objekte bei beiden Partitionen in derselben Gruppe liegen oder sich bei beiden in verschiedenen Klassen befinden. Der Randindex kann Werte zwischen *0* und *+1* annehmen, je größer diese Zahl ist, desto größer ist die Übereinstimmung der Zuordnung. Eine Weiterentwicklung ist der *adjusted Randindex*, der zusätzlich eine Korrektur in Bezug auf den Zufall in die Berechnung aufnimmt (vgl. Öllinger 2010, S. 44f.).

Bei einer Clusterung mit drei Clustern ist die Übereinstimmung zwischen *k-means*-Verfahren der Rohdaten und dem *PAM*-Verfahren am größten (siehe Tabelle 4.3.1). Am meisten unterscheidet sich die Partition, die sich aus dem *k-means*-Verfahren der Splinekoeffizienten ergibt, von den beiden anderen.

Betrachtet man dagegen eine Clusterung mit vier Klassen, so weicht das *PAM*-Verfahren von den Zuordnungen des *k-means-Verfahren* ab. So wird Graz West aus dem Cluster mit den höchsten Werten entfernt und dem mittleren Cluster zugeteilt. Knittelfeld und Zeltweg wandern in das Cluster mit den niedrigsten Werten und Fürstenfeld und Köflach kommen in die vierte Gruppe.

	3 Cluster		4 Cluster	
	k-means (Rohdaten)	*PAM*	*k-means (Rohdaten)*	*PAM*
k-means (Spline)	0.3	0.22	0.74	0.44
k-means (Rohdaten)		0.64		0.49

Tab. 4.3.1. PM_{10}: Randindex für den Vergleich der drei Cluster-Methoden

Um eine gute räumliche Vorstellung der Cluster zu erhalten, werden die einzelnen Messstationen und ihre Zuordnung zu den Klassen in der Steiermark Karte eingezeichnet (siehe Abbildung 4.3.9). Objekte, die im selben Cluster liegen, bekommen dieselbe Farbe, wobei die Klasse mit den höchsten Werten mit Rot, jene mit den mittleren[19]

[18] In der Statistik Software *R* ist der *Randindex* als randIndex bzw. als *adjustedRandIndex* (package *flexclust*) implementiert.

[19] Ein „mittleres" Cluster ist nicht immer eindeutig zu den beiden anderen abgegrenzt, bei der Kennzeichnung wird trotzdem eine Unterscheidung versucht.

Werten mit Grün und jene mit den niedrigsten Werten mit Blau markiert werden. Zusätzlich wird für jede Clusteranalysemethode ein eigenes Symbol (Kreisfläche, Stern, Dreieck) verwendet, wobei die Zuordnung nach dem *PAM*-Verfahren als Ausgangspunkt gewählt wird. Wenn sich die Einteilung bei einem Verfahren von dieser Standardeinteilung unterscheidet, wird ein zusätzliches Symbol in der entsprechenden Farbe eingefügt.

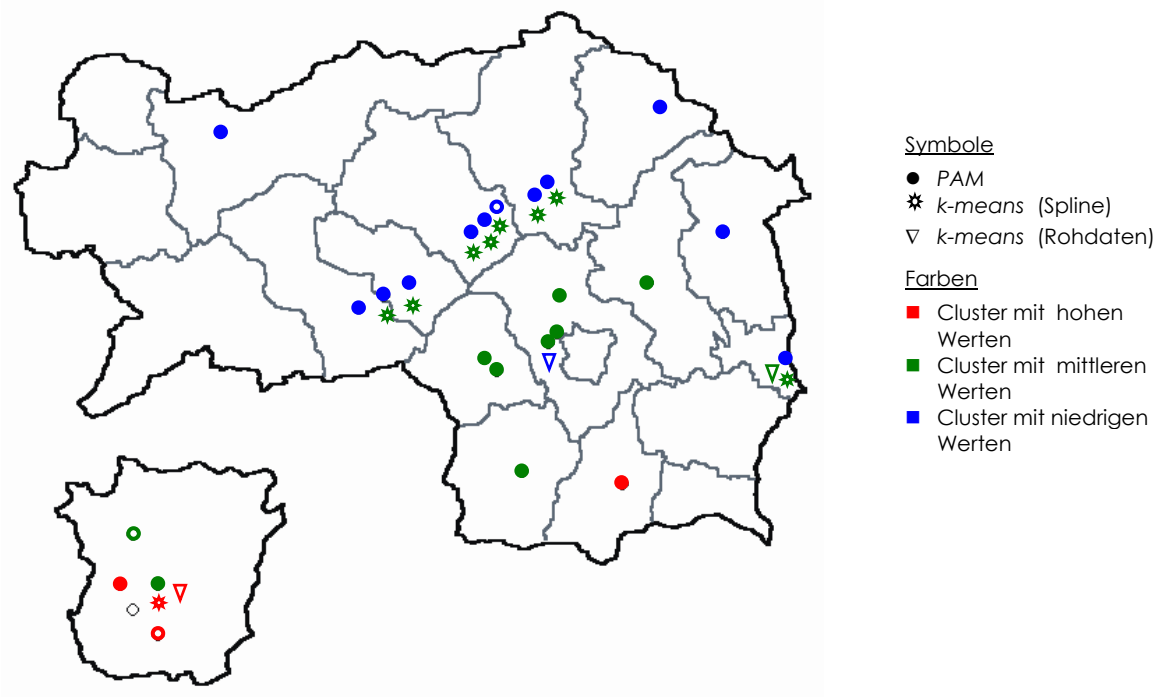


Abb. 4.3.9. PM_{10} – Cluster in der Steiermark. Repräsentanten[20] nach dem PAM-Verfahren sind Niklasdorf, Graz Nord und Graz Süd.

Bei dieser Darstellung erkennt man, dass das *PAM*-Verfahren stark „räumlich" clustert, d.h. Objekte werden rund um die Medoids gebildet. Danach kann man die Steiermark in einen nord-östlichen (niedrigere Werte), mittleren und südlicheren (hohe Werte) Bereich aufteilen. Beim *k-means*-Verfahren der Splinekoeffizienten ist die Gruppe der mittleren Werte[21] sehr groß und umfasst die Stationen nördlich von Graz. In dem Cluster mit den niedrigen Werten befinden sich nur vier Stationen, die eher in Randgebieten der Steiermark liegen (Liezen, Mürzzuschlag, Masenberg und Fürstenfeld). Dagegen befinden sich die Stationen mit den höchsten Werten, die Grazer Stationen und Leibnitz, im südlichen Teil der Steiermark und eher in „kesselartigen" Gebieten (geringe Seehöhe).

[20] Die Repräsentanten sind als Kreise ohne Füllung markiert.
[21] Symbol: ✸

Eine nächste Überlegung ergibt sich bei der Frage, wie sich die Cluster mit der Wahl der Knoten für die Erzeugung der Spline-Funktionen verändern. Auch hier bietet sich der Randindex als eine Kennzahl der Übereinstimmungen bzw. Nichtübereinstimmungen für einen Vergleich an. Im Folgenden reicht es, den Randindex für die Methoden zu berechnen, die die geschätzten Splinekoeffizienten clustern. In Tabelle 4.3.2 befinden sich die Randindices, wenn Splinekoeffizienten mithilfe von 18, 36 bzw. 72 inneren Knoten erzeugt werden. Für das *k-means*-Verfahren gibt es keinen Unterschied, ob man 18 oder 36 innere Knoten wählt. Bei einer höheren Anzahl an Knoten verändern sich die Cluster sehr stark. Der Grund dafür kann darin liegen, dass mehr Informationen für die Clusterung vorhanden sind. Der vorliegende Datensatz ist jedoch in Hinblick auf fehlende Werte bereinigt, indem diese durch den jeweiligen Stationsmittelwert ersetzt wurden. Dadurch verändert sich natürlich auch der Verlauf der Funktionen, d.h. es gibt konstante Abschnitte. Das *PAM*-Verfahren dagegen scheint mit Erhöhung der Anzahl der Knoten ähnliche Cluster zu erzeugen.

	k-means (Spline)		*PAM*	
	18 Knoten	*72 Knoten*	*18 Knoten*	*72 Knoten*
36 Knoten	1	0.33	0.22	0.76
18 Knoten		0.33		0.17

Tab. 4.3.2. PM_{10}: Randindices bei einer Partition von drei Clustern, mit unterschiedlicher Anzahl an Knoten beim Erzeugen der funktionalen Daten

3.2 Ergebnisse zu Ozon

Nach dem Screeplot ergibt sich für den Luftschadstoff Ozon eine optimale Aufteilung in drei Klassen. Da insgesamt nur zehn der ausgewählten Messstationen diesen Schadstoff messen, sind die Gruppen sehr klein, und zwar mit den Clustergrößen 3, 1 und 6. Beide *k-means*-Verfahren, und zwar das Clustern der geschätzten Splinekoeffizienten und das Clustern der Rohdaten, führen zu denselben drei Gruppierungen. Eine Klasse enthält nur eine Messstation, und zwar Masenberg (siehe Abbildung 4.3.10 und 4.3.12).

Die Messwerte der Messstelle Masenberg liegen deutlich über denjenigen der anderen. Die Funktionen der übrigen Stationen weisen nur geringfügige Unterschiede auf, sowohl in den Ausprägungen als auch im Verlauf. Das bedeutet, dass die beiden Clusterzentren sehr nahe beieinander liegen. Möglicherweise ist, trotz Entscheidung mithilfe des Screeplots, eine Aufteilung in drei Gruppen weniger sinnvoll als eine in zwei. Damit ergeben sich zwei Klassen, wobei die Gruppe mit den hohen Werten die Messstation Masenberg enthält und die zweite Gruppe die restlichen neun (siehe Abbildung 4.3.11).

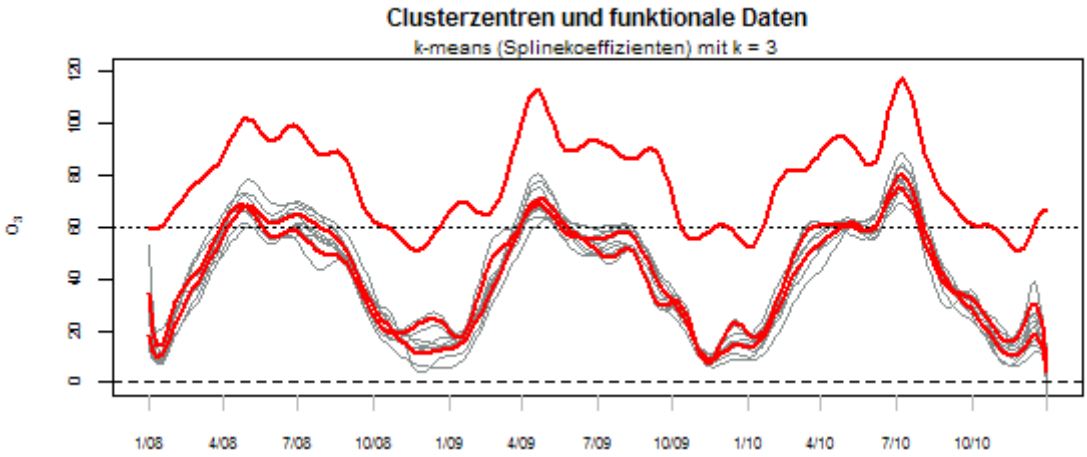


Abb. 4.3.10. O_3 – Clusterzentren (rote Funktionen) mit den funktionalen Daten der Messstationen bei einer Clusterung der geschätzten Splinekoeffizienten (*k-means*, *k* = 3)

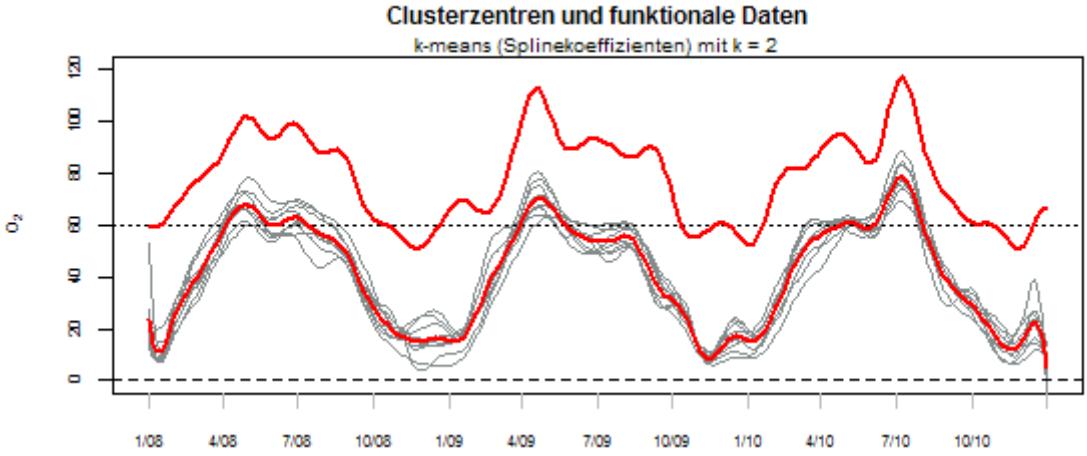


Abb. 4.3.11. O_3 – Clusterzentren (rote Funktionen) mit den funktionalen Daten der Messstationen bei einer Clusterung der geschätzten Splinekoeffizienten (*k-means*, *k* = 2)

Beim *PAM* ergibt sich für den Luftschadstoff Ozon eine optimale Repräsentantenanzahl von zwei. Die durchschnittliche Silhouette-Breite liegt bei dieser Clusterung bei 0.67, jedoch ergibt sich für das zweite Cluster eine durchschnittliche Silhouette-Breite von 0, d.h. es gibt eine Gruppe mit nur einem Objekt. Bei drei Gruppierungen liegt die durchschnittliche Silhouette-Breite nur mehr bei 0.22 und es gibt wiederum eine Gruppe mit einer Messstation (Masenberg). Die Zuordnung zu den drei Gruppen ist dieselbe wie bei den anderen beiden Methoden. Die Repräsentanten sind nun **Liezen**, **Masenberg** und **Weiz**.

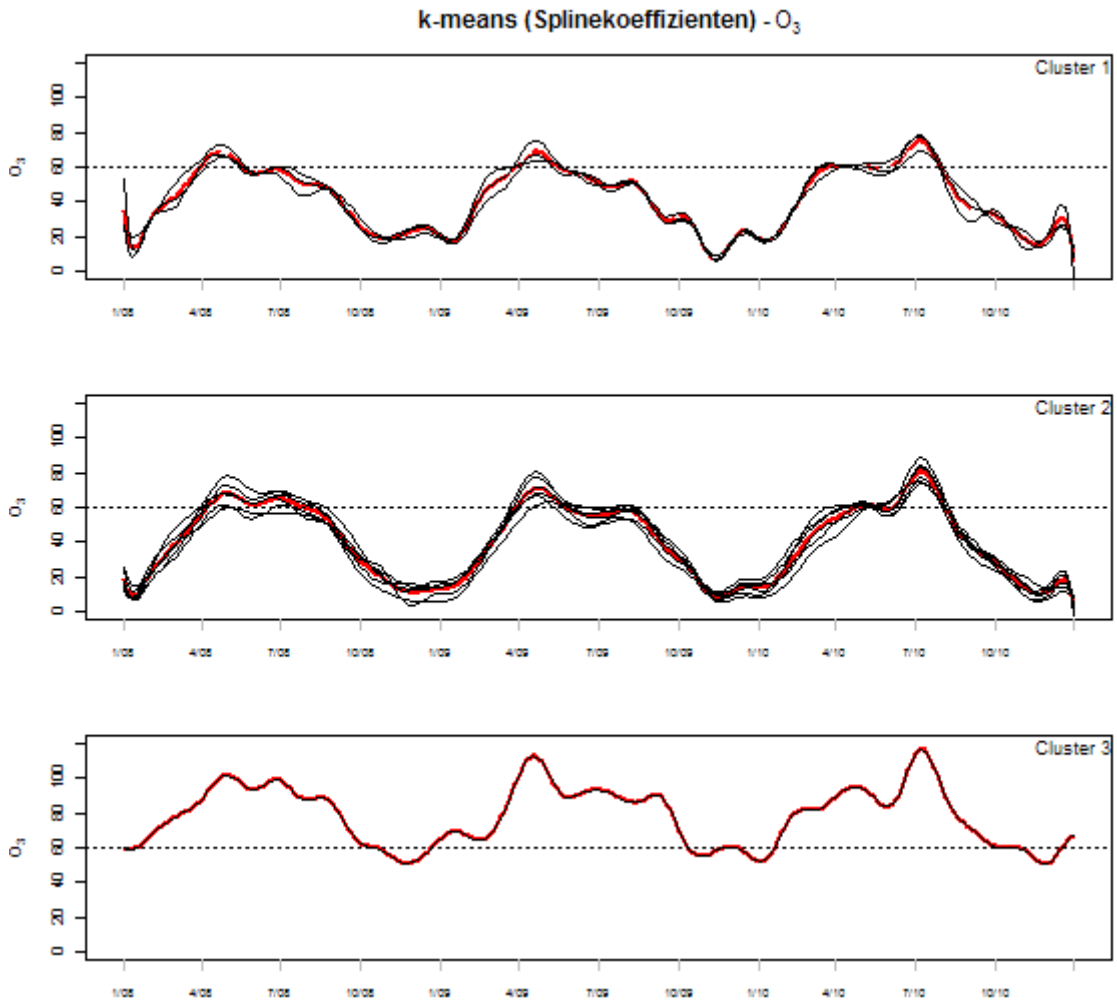


Abb. 4.3.12. O_3 – Cluster beim k-means der geschätzten Splinekoeffizienten (rote Funktion: Clusterzentrum)
Cluster 1: Judenburg, Liezen, Mürzzuschlag
Cluster 2: Deutschlandsberg, Fürstenfeld, Voitsberg, Weiz, Graz Nord, Graz Süd
Cluster 3: Masenberg

Für eine Clusterung mit drei Clustern sind die Randindices natürlich gleich eins, da die Übereinstimmung exakt ist (siehe Tabelle 4.3.3). Bei einer höheren Anzahl von Klassen ergeben sich sehr wohl Unterschiede bei der Zuordnung der Objekte. Dabei erzeugen das *PAM*-Verfahren und *k-means*-Verfahren der Splinekoeffizienten dieselben Cluster, während es weniger Übereinstimmungen zu der Clusterung der Rohdaten gibt. Das *k-means*-Verfahren scheint für diese Daten robust in Hinblick der Knotenwahl zu sein, es werden jeweils dieselben Cluster erzeugt[22]. Bei der *PAM*-Methode liefern 36 bzw. 72 innere Knoten dieselben Gruppierungen, weniger Knoten führen zu einer anderen Zuordnung.

[22] Die Tabelle mit den Randindices befindet sich in Anhang A.4.

	3 Cluster		*4 Cluster*	
	k-means (Rohdaten)	*PAM*	*k-means (Rohdaten)*	*PAM*
k-means (Spline)	1	1	0.41	1
k-means (Rohdaten)		1		0.41

Tab. 4.3.3. O_3: Randindex für den Vergleich der drei Cluster-Methoden

Räumlich gesehen gibt es niedrigere Werte in der Obersteiermark und mittlere im Süden der Steiermark. Auffallend ist die Messstation Masenberg mit deutlich höheren Werten als die übrigen.

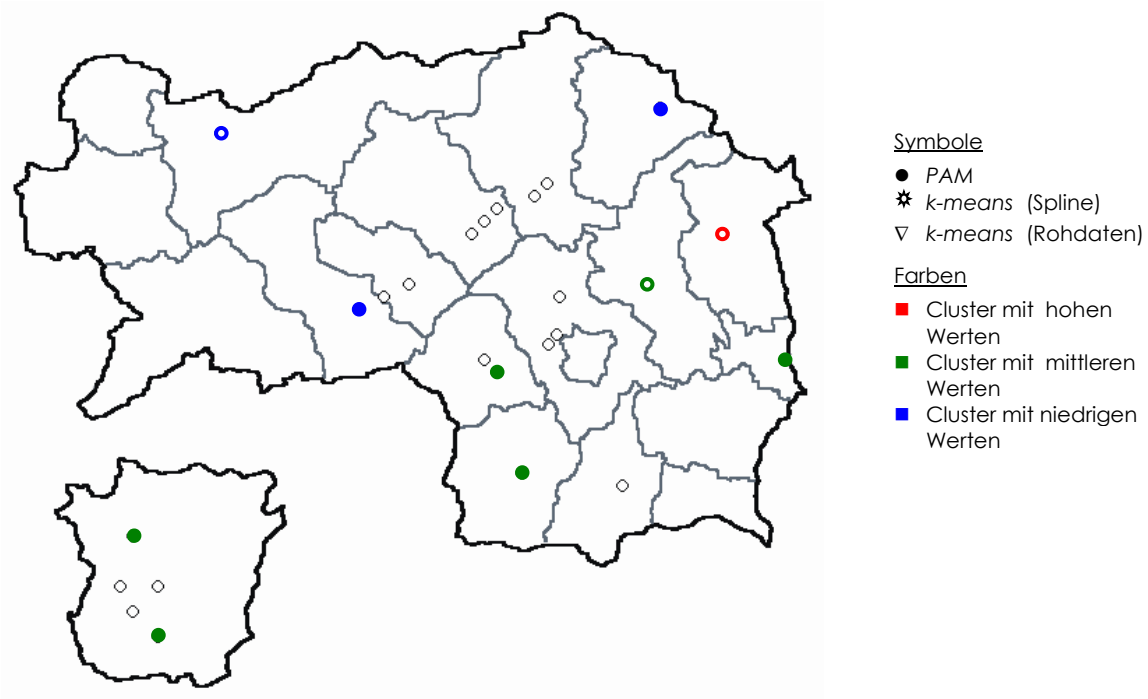


Abb. 4.3.13. O_3 – Cluster in der Steiermark. Repräsentanten[23] nach dem PAM-Verfahren sind Liezen, Weiz und Masenberg.

3.3 Ergebnisse zum Luftschadstoff Schwefeldioxid

Nach dem Screeplot ist für das *k-means*-Verfahren (sowohl für das Clustern der Splinekoeffizienten als auch der Rohdaten) eine Aufteilung in drei Gruppen optimal, hier ist der Sprung zur kleineren Klassenanzahl sehr stark ausgeprägt. Diese Clusterung erzeugt eine Klasse mit nur einer Messstelle, eine Gruppe mit zwei Stationen und die letzte beinhaltet die restlichen 14 (siehe Abbildungen 4.3.14 und 4.3.15). Die Zuordnung der Messstationen zu den einzelnen Clustern ist sowohl für das Clustern der (geschätzten) Splinekoeffizienten als auch der Rohdaten dieselbe.

[23] Die Repräsentanten sind als Kreise ohne Füllung markiert.

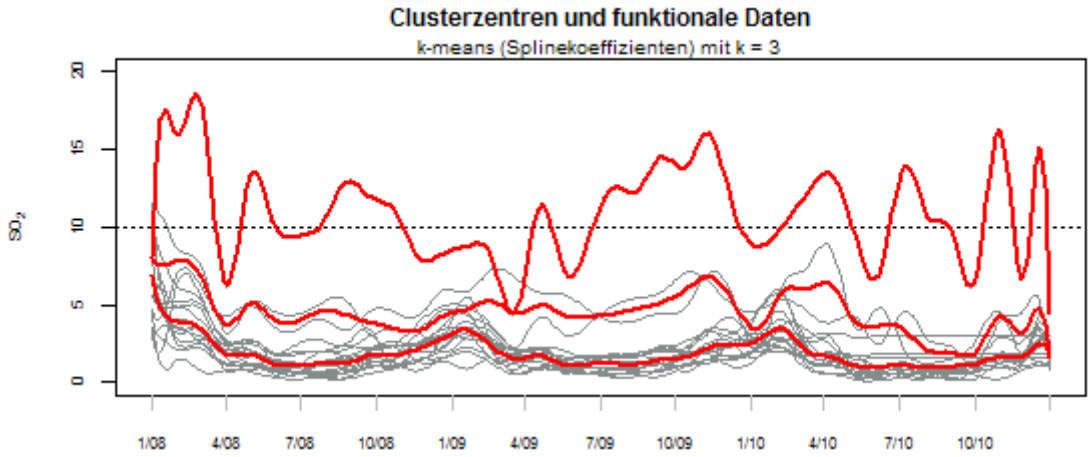


Abb. 4.3.14. SO_2 – Clusterzentren (rote Funktionen) mit den funktionalen Daten der Messstationen bei einer Clusterung der geschätzten Splinekoeffizienten (*k-means*, *k = 3*).

Ein Cluster enthält als einziges Objekt die Messstation Straßengel. Deren Funktion ist im Verlauf mit sehr ausgeprägten (lokalen) Extrema auffallend, sie ähnelt keiner der anderen Funktionen. Vergleicht man die Messwerte dieser Messstation mit jenen der anderen, wird diese außergewöhnliche Form der Funktion klar[24]. Die meisten der anderen Messstellen haben Messwerte, die zwischen 0 und 7 [µg/m³] liegen, wobei es höhere Werte um Neujahr gibt. In Straßengel jedoch sind die beobachteten Daten „gleichmäßiger" verteilt und lassen weniger saisonale Schwankungen erkennen. Hierbei stellt sich die Frage, welche (zusätzlichen) Einflussfaktoren diese Werte ergeben. Im dritten Cluster befinden sich die Messstationen Judendorf, das nahe Straßengel liegt, und Leoben/Donawitz. Auch deren Funktionen weichen von den anderen ab, jedoch nicht so stark wie jene von Straßengel. Bei der Zuordnung der Objekte zu den Clustern kommen diese funktionalen Charakteristika zum Tragen.

Für die Clusterung nach dem *PAM*-Verfahren erhält man die höchste durchschnittliche Silhouette-Breite für zwei Cluster, jedoch ist die Silhouette-Breite auch noch für drei Gruppen recht hoch (0.75 bei zwei Gruppen, 0.46 bei drei Gruppen). Dabei gibt es eine ein-elementige Menge (mit der Messstation Straßengel) und die übrigen bilden das zweite Cluster (siehe Abbildung 4.3.16). Bei einer Aufteilung in drei Gruppen kommt man zu derselben Clusterung wie mit den vorherigen Methoden. Die Repräsentanten sind die Messorte **Straßengel**, **Leoben/Donawitz** und **Voitsberg**. Auch bei einer höheren Anzahl ergeben sich dieselben Gruppierungen.

[24] Die Graphen der Funktionen für jede Messstation, gemeinsam mit den Messwerten geplottet, befinden sich in Anhang B.1.

Die Zuordnung beim *k-means*-Verfahren bleibt dieselbe, auch wenn die Funktionen mit einer anderen Knotenmenge approximiert werden. Beim *PAM* ist die Partition ident für die Funktionen mit 18 und 72 inneren Knoten. Die Übereinstimmung bei einer Clusterung der Splines mit 36 inneren Knoten ist relativ hoch[25].

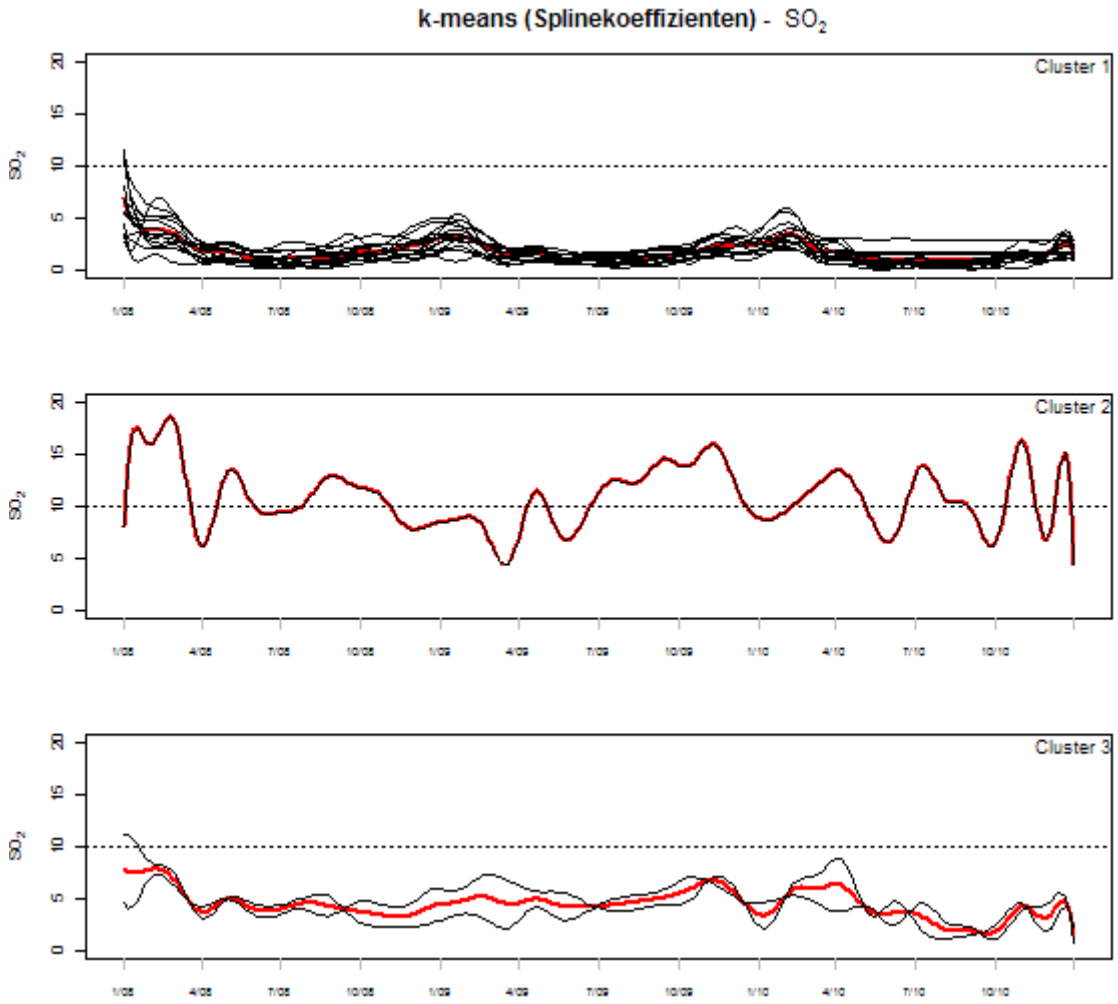


Abb. 4.3.15. SO_2 – Cluster mit den funktionalen Daten der Messstationen (rote Funktion: Clusterzentrum)
Cluster 1: Bruck, Deutschlandsberg, Fürstenfeld, Knittelfeld, Köflach, Leoben/Göß, Liezen, Masenberg, Niklasdorf, Peggau, Voitsberg, Graz Nord, Graz Süd, Graz West
Cluster 2: Straßengel
Cluster 3: Judendorf, Leoben/Donawitz

[25] Die Tabelle mit den Randinidices befindet sich in Anhang A.4.

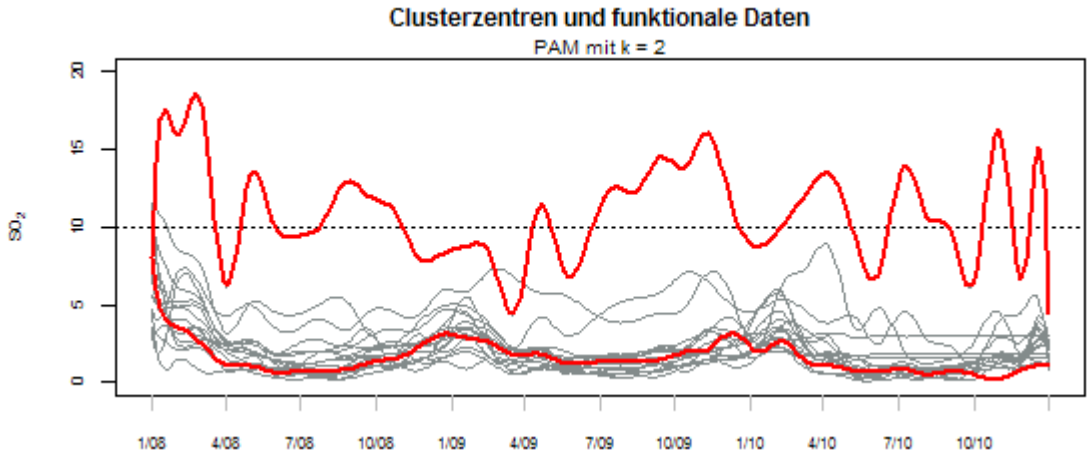


Abb. 4.3.16. SO_2 – Clusterzentren (rote Funktionen) mit den funktionalen Daten der Messstationen bei einer Clusterung der geschätzten Splinekoeffizienten (*PAM*, *k* = 2).

Drei Messstationen im zweiten und im dritten Cluster sind auffallend. Zwei davon, Straßengel und Judendorf, sind räumlich sehr nahe, Leoben/Donawitz ist nördlich davon (siehe Abbildung 4.3.17). Hier scheint es (weitere) Einflussfaktoren zu geben, die sich nicht auf die anderen Messstationen auswirken. Man bedenke, dass im Umfeld dieser Messstellen die Papierfabrik Sappi bzw. die Stahlwerke Donawitz liegen. Interessant ist die Verteilung der Messorte aus dem ersten Cluster. Diese liegen über die gesamte Steiermark verteilt, hier kann man kaum von einem kugelförmigen Cluster sprechen.

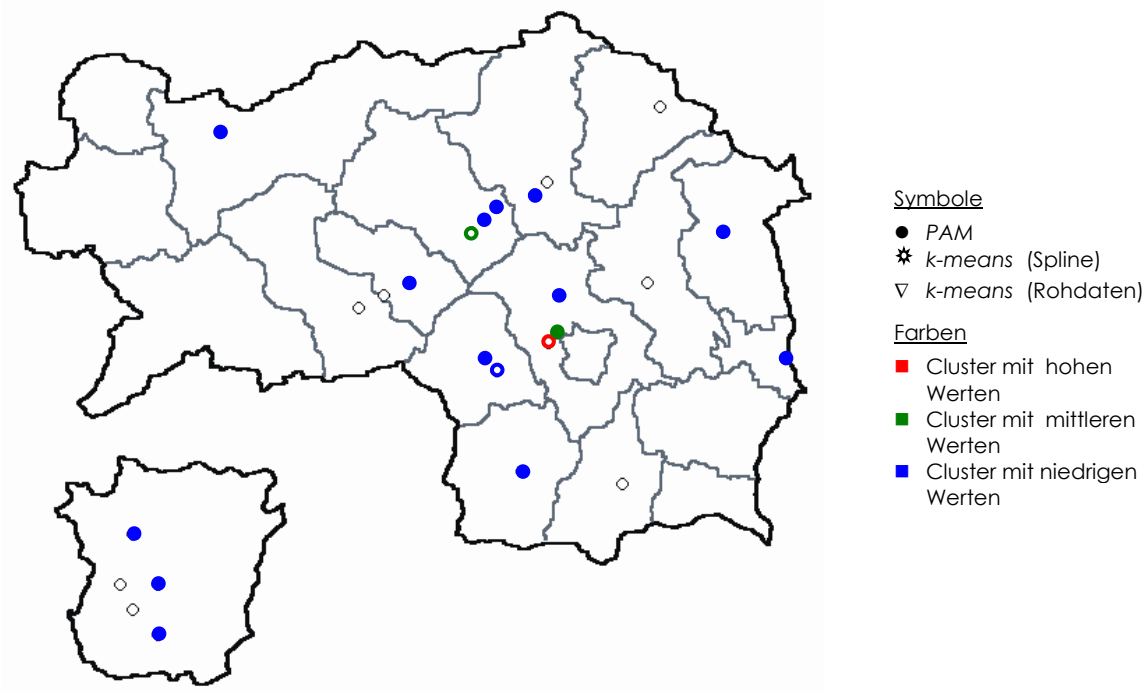


Abb. 4.3.17. SO_2 – Cluster in der Steiermark. Repräsentanten nach dem PAM-Verfahren sind Voitsberg, Leoben/Donawitz und Straßengel.

3.4 Ergebnisse zum Luftschadstoff Stickstoffdioxid

Für die Clusterung mit dem *k-means*-Verfahren ist eine Aufteilung in drei Gruppen optimal. Bei beiden *k-means*-Verfahren wird jeweils eine Gruppe mit nur einem Objekt erzeugt, wobei einmal die Messstation mit den niedrigsten Werten (*k-means* nach geschätzten Splinekoeffizienten) und das andere Mal jene mit den höchsten Werten (*k-means* nach Rohdaten) enthalten ist (siehe Abbildung 4.3.18). Die beiden anderen Cluster ergeben sich dann aufgrund der Wahl dieses Clusters, in je eine Gruppe mit mittleren und hohen Werten bzw. in je eine mit mittleren und niedrigen Werten.

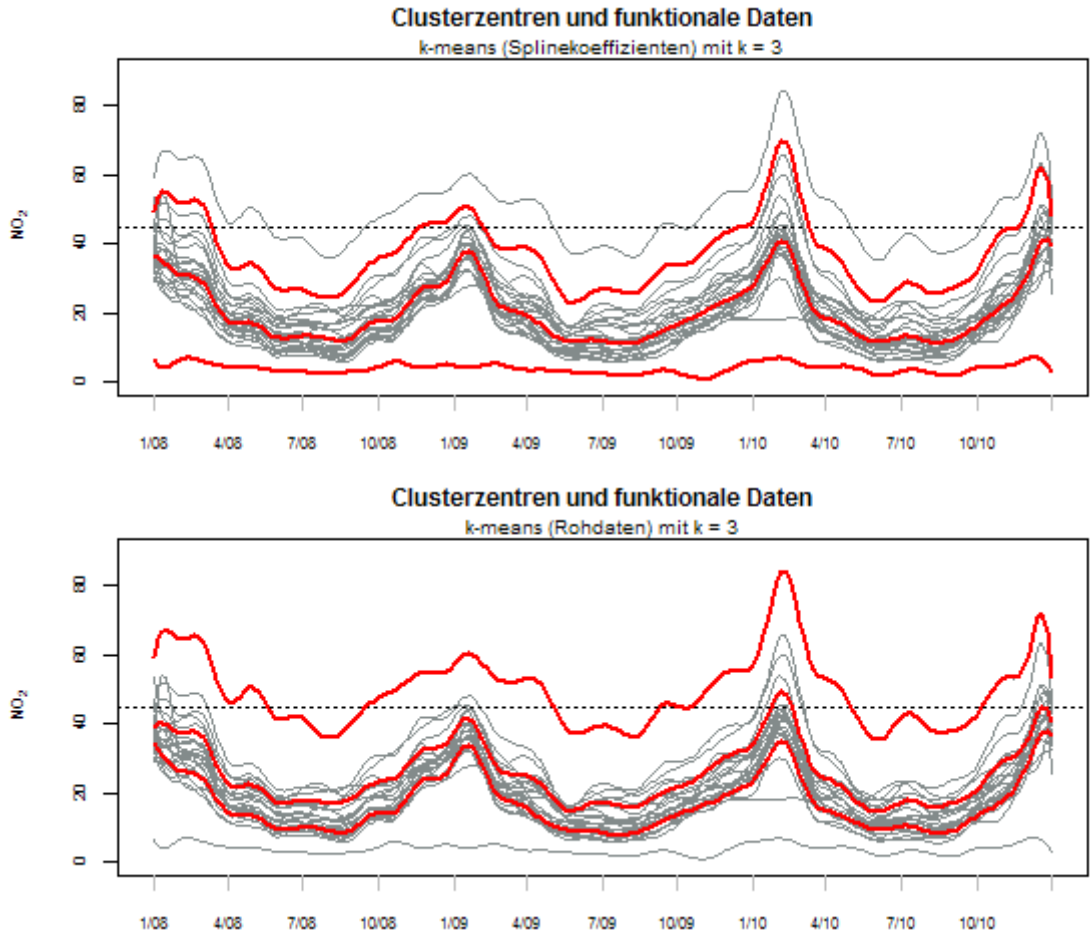


Abb. 4.3.18. NO_2 – Clusterzentren (rote Funktionen) mit den funktionalen Daten der Messstationen bei einer Clusterung der geschätzten Splinekoeffizienten, *k-means, k = 3* (obere Graphik) bzw. bei einer Clusterung der Rohdaten, *k-means*, *k = 3* (untere Graphik).

Für das *k-means*-Verfahren der Splinekoeffizienten werden drei Grazer Messstationen, die die höchsten Werte messen, zu einem Cluster zusammengefasst. Damit ergibt sich eine große Gruppe mit mittleren Werten, die insgesamt 20 Messstellen beinhaltet. Beim *k-means*-Verfahren nach den Rohdaten unterscheiden sich die beiden Klassen der

niedrigen und mittleren Werten nach der Anzahl der Objekte nicht so stark. Die mittlere Klasse enthält 13 Messstationen und die mit den niedrigen Werten 10.

Auffallend sind zwei Kurven: jene, deren Werte nahe bei null sind (Masenberg), und jene, die deutlich über allen anderen liegt (Graz Don Bosco). Zudem hat die Funktion der Messstation Masenberg einen ungewöhnlichen Verlauf im Vergleich zu den anderen. Deren Werte sind beinahe konstant über den gesamten Zeitraum hinweg, d.h. es sind keine Tendenzen erkennbar, weder saisonal noch ein Gesamttrend. Dagegen ist der Verlauf der Funktion der Messstelle Graz Don Bosco ähnlich zu den anderen, auffallend ist sie nur aufgrund der extrem hohen Werte. Grundsätzlich sind die Messwerte der Grazer Stationen, vor allem Graz Don Bosco, Graz Süd und Graz West, höher als die der Übrigen.

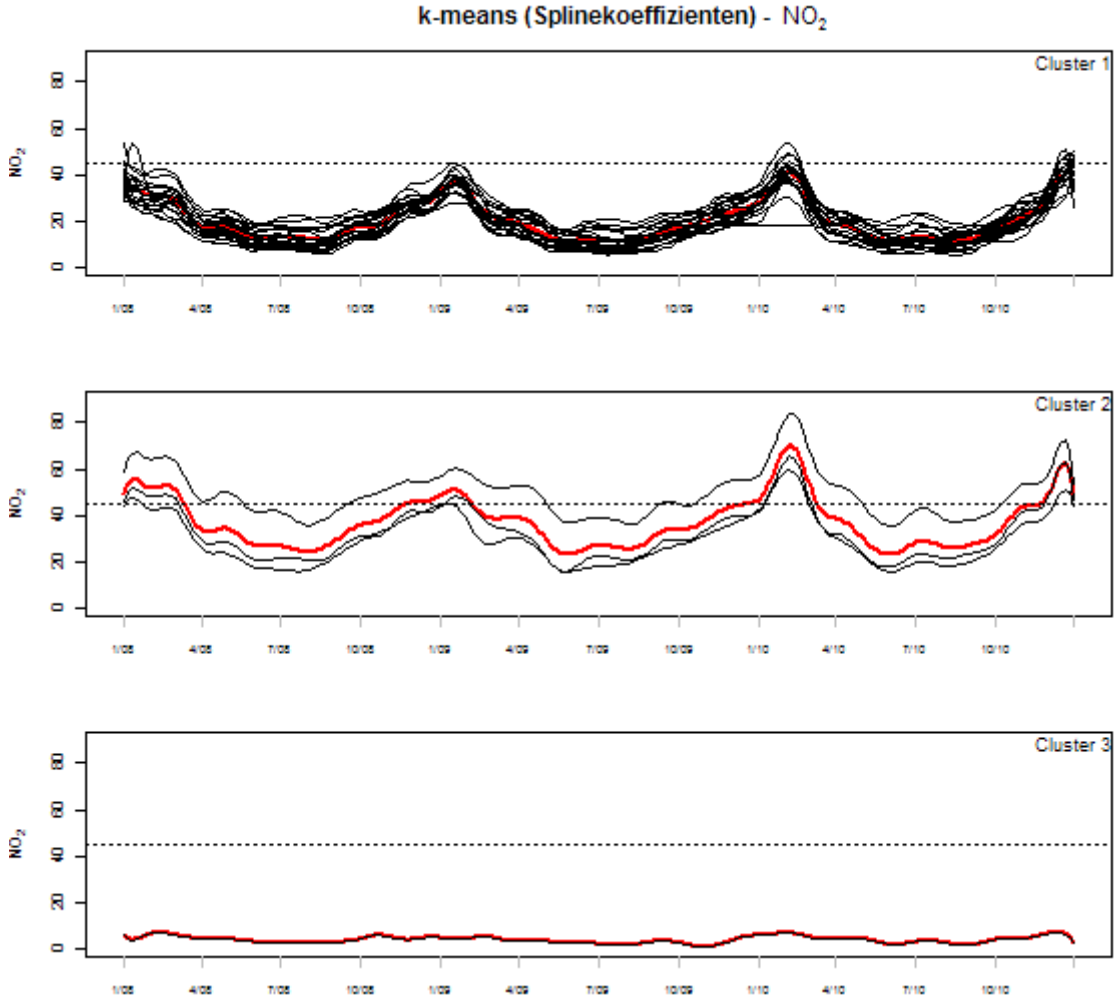


Abb. 4.3.19. NO_2 – Cluster, bei Clustern der geschätzten Splinekoeffizienten
Cluster 1: Bruck, Deutschlandsberg, Fürstenfeld, Judenburg, Judendorf, Kapfenberg, Knittelfeld, Köflach, Leibnitz, Leoben/Donawitz, Leoben/Göß, Liezen, Mürzzuschlag, Niklasdorf, Peggau, Straßengel, Voitsberg, Weiz, Zeltweg, Graz Nord
Cluster 2: Graz Don Bosco, Graz Süd, Graz West
Cluster 3: Masenberg

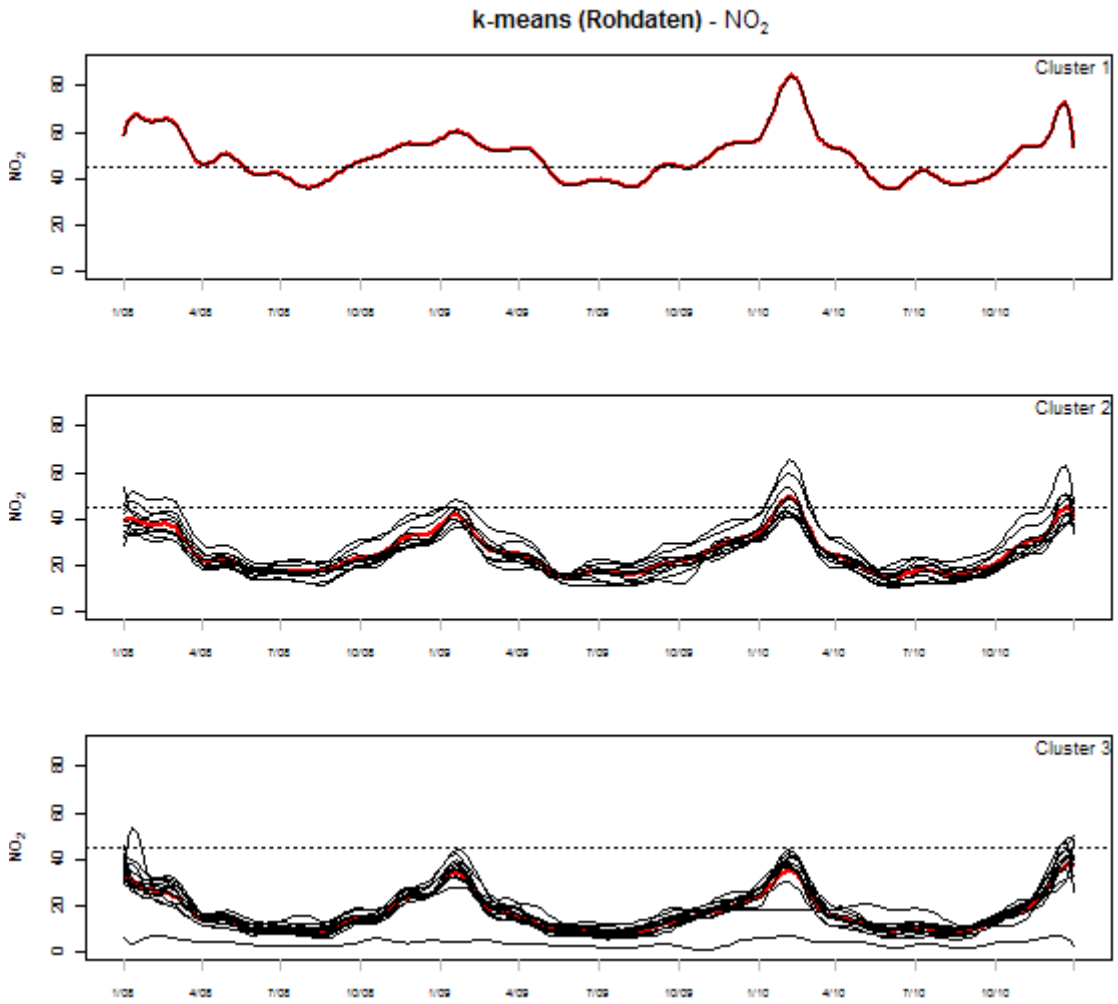


Abb. 4.3.20. NO_2 – Cluster, bei Clustern der Rohdaten
Cluster 1: Graz Don Bosco,
Cluster 2: Judendorf, Köflach, Leibnitz, Leoben/Göß, Peggau, Straßengel, Weiz, Graz Nord, Graz Süd, Graz West
Cluster 3: Bruck, Deutschlandsberg, Fürstenfeld, Judenburg, Kapfenberg, Knittelfeld, Leoben/Donawitz, Liezen, Masenberg, Mürzzuschlag, Niklasdorf, Voitsberg, Zeltweg

Bei einer Clusterung mit mehr Klassen, z. B. $k = 4$, bilden die beiden Messstationen Graz Don Bosco und Masenberg jeweils ein Cluster. Die übrigen Stationen werden auf die beiden mittleren Klassen aufgeteilt (Cluster 2: Judendorf, Köflach, Leibnitz, Leoben/Göß, Peggau, Straßengel, Weiz, Graz Nord, Graz Süd, Graz West; Cluster 3: Bruck, Deutschlandsberg, Fürstenfeld, Judenburg, Kapfenberg, Knittelfeld, Leoben/Donawitz, Liezen, Mürzzuschlag, Niklasdorf, Voitsberg, Zeltweg). Damit ergeben sich zwei ein-elementige Gruppen, eine mit 10 und eine mit 12 Objekten.

Wenn man diese Zuordnung mit derjenigen vergleicht, die sich aus dem *k-means*-Verfahren nach den Rohdaten für drei Cluster ergibt, erkennt man Ähnlichkeiten. So stellt auch hier die Messstation Graz Don Bosco eine eigene Klasse dar. Cluster 2 und 3 enthalten bis auf Masenberg dieselben Objekte wie beim Clustern mit vier Gruppierungen

(siehe Abbildung 4.3.20). Das Verfahren unterscheidet stärker bei den Funktionen mit mittleren Werten, wenn die Funktion mit den höchsten Werten bereits ausgeschlossen wurde.

Beim *Partitioning Around Medoids*-Verfahren ergibt sich mithilfe der Informationen der Silhouette-Breiten, dass eine Aufteilung in vier Klassen am besten ist. Die durchschnittliche Silhouette-Breite liegt bei 0.254 (siehe Abbildung 4.3.21,). Dabei ergibt sich dieselbe Zuordnung der Objekte wie beim *k-means*-Verfahren mit vier Klassen (siehe Abbildung 4.3.23). Die Medoids sind **Bruck**, **Graz Nord**, **Masenberg** und **Graz Don Bosco**. Dabei erhält man aber zwei Klassen mit jeweils einem Element, und zwar mit Masenberg und Graz Don Bosco (Silhouette-Breiten der Klassen: 0.358, 0.195, 0.0, 0.0). Zudem sind die Zuordnungen zu einem Cluster für sieben Messstationen nicht eindeutig ($|s(i)| < 0.2$). Fürstenfeld, Graz West, Judendorf, Leibnitz, Peggau, Köflach und Leoben/Donawitz liegen zwischen der ersten und zweiten Gruppe.

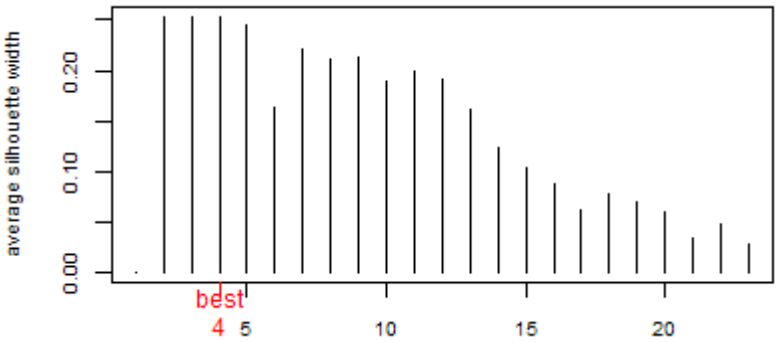


Abb. 4.3.21. NO_2: durchschnittliche Silhouette-Breiten für unterschiedliche Anzahl an Repräsentanten

Die nächstbeste durchschnittliche Silhouette-Breite unterscheidet sich nur geringfügig von der besten (0.253). Diese wird bei einer Klassenanzahl von $k = 3$ erzielt. Auch die durchschnittlichen Silhouette-Breiten der einzelnen Klassen sind in der Größenordnung der Clusterung mit vier Gruppen (Silhouette-Breiten der Klassen: 0.248, 0.282, 0.0). Bei dieser Klassenanzahl befinden sich fünf Messstationen zwischen Cluster 1 und 2 ($|s(i)| < 0.2$), und zwar Kapfenberg, Mürzzuschlag, Fürstenfeld, Köflach und Leoben/Donawitz.

Für einen Vergleich mit den anderen Verfahren ist eine Partition mit drei Klassen gut geeignet. Diese führt zu derselben Struktur wie beim *k-means*-Verfahren der Rohdaten mit drei Cluster (siehe Abbildung 4.3.22).

Für den Luftschadstoff Stickstoffdioxid ist entscheidend, welches Verfahren man bei einer Clusteranzahl $k = 3$ wählt (siehe Tabelle 4.3.4). Eine große Übereinstimmung bei den Clusterzuordnungen gibt es nur zwischen dem *PAM*-Verfahren und dem *k-means*-Verfahren der Rohdaten. Das *k-means*-Verfahren der Splinekoeffizienten unterscheidet sich hingegen sehr stark von den beiden anderen Verfahren.

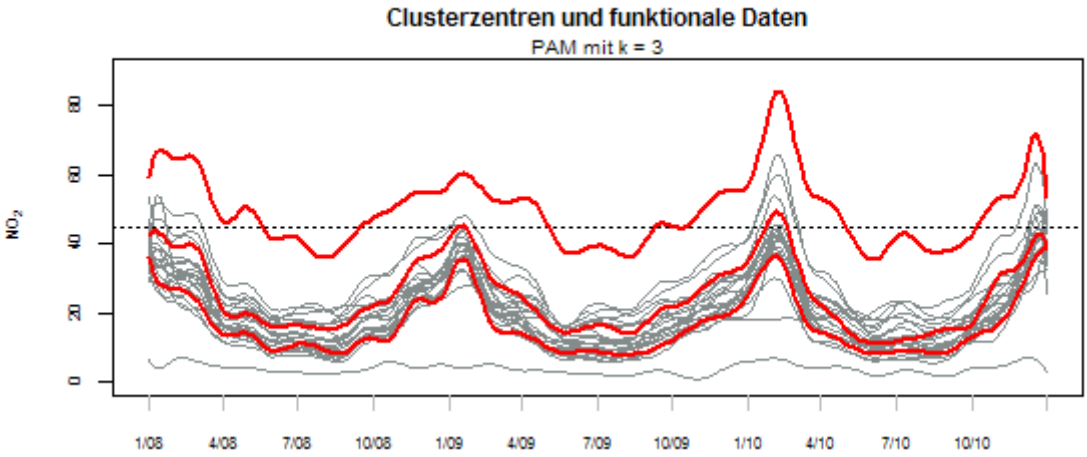


Abb. 4.3.22. NO_2 – Clusterzentren mit den funktionalen Daten, *PAM*-Verfahren

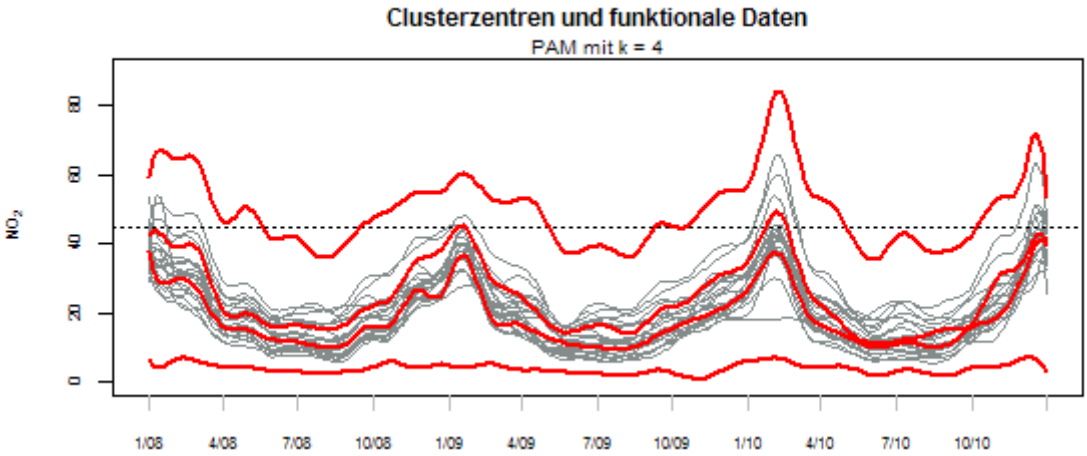


Abb. 4.3.23. NO_2 – Cluster, bei Clustern mit *PAM* und vier Klassen
Cluster 1: Bruck, Deutschlandsberg, Fürstenfeld, Judenburg, Kapfenberg, Knittelfeld, Liezen, Masenberg, Mürzzuschlag, Niklasdorf, Voitsberg, Zeltweg,
Cluster 2: Judendorf, Köflach, Leibnitz, Leoben/Donawitz, Leoben/Göß, Peggau, Straßengel, Weiz, Graz Nord, Graz Süd, Graz West
Cluster 3: Masenberg,
Cluster 4: Graz Don Bosco

Tabelle 4.3.5 zeigt, dass für diese Daten das *PAM*-Verfahren weniger stark auf die Knotenanzahl reagiert. Für eine kleine und für eine große Anzahl an Knoten werden dieselben drei Cluster erzeugt. Für 36 innere Knoten verändert sich die Zuordnung der Messstationen. Dafür gibt es starke Abweichungen beim *k-means*-Verfahren, und zwar wenn mehr Knoten gewählt werden. Auch hier könnte ein Grund darin liegen, dass die fehlenden Werte durch einen konstanten Wert, dem jeweiligen Stationsmittelwert, ersetzt wurden.

	3 Cluster		*4 Cluster*	
	k-means (Rohdaten)	*PAM*	*k-means (Rohdaten)*	*PAM*
k-means (Spline)	0.12	0.1	1	0.84
k-means (Rohdaten)		0.84		0.84

Tab. 4.3.4. NO_2: Randindex für den Vergleich der drei Cluster-Methoden

	k-means		*PAM*	
	18 Knoten	72 Knoten	18 Knoten	72 Knoten
36 Knoten	1	0.12	0.69	0.69
18 Knoten		0.12		1

Tab. 4.3.5. NO_2: Randindices bei einer Partition von drei Cluster, mit unterschiedlicher Anzahl an Knoten beim Erzeugen der funktionalen Daten

Anhand der Steiermark Karte erkennt man, dass die Luftverschmutzung durch Stickstoffdioxid in der Stadt Graz und nahe Graz stärker ausgeprägt ist als im restlichen Land (siehe Abbildung 4.3.24). Vor allem im Norden und im Osten findet man eher niedrigere Werte. Das *k-means*-Verfahren (der Splinekoeffizienten) unterscheidet noch weniger zwischen den einzelnen Messstationen, da wird eine große Gruppe von Stationen, die über die gesamte steirische Fläche verteilt sind, zu einem Cluster zusammengefasst. Nur die Grazer Messstellen und Masenberg werden eigenen Clustern zugeordnet.

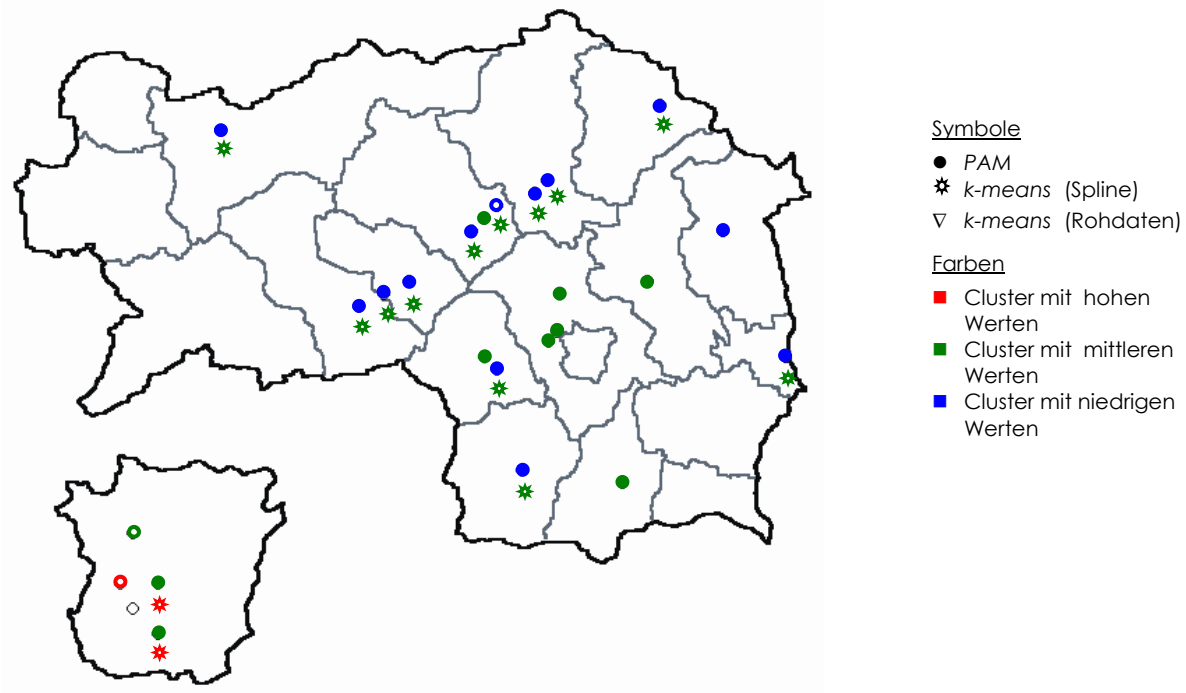


Abb. 4.3.24. NO_2 – Cluster in der Steiermark. Repräsentanten nach dem PAM-Verfahren sind Niklasdorf, Graz Nord und Graz Don Bosco.

V

SCHLUSSBEMERKUNGEN

In der vorliegenden Untersuchung wurden 24 (der 25 ausgewählten) Messstationen der Steiermark in Bezug auf die Luftschadstoffe Feinstaub PM_{10}, Ozon O_3, Schwefeldioxid SO_2 und Stickstoffdioxid NO_2 untersucht, wobei sich die Analyse auf Daten aus drei Jahren (1096 Tage) bezieht. Die Daten liegen in Form von Tagesmittelwerten vor, jedoch liefert nicht jede Messstation für alle Merkmale und für jeden Tag Werte. Fehlende Werte wurden durch den jeweiligen Stationsmittelwert ersetzt, solange mehr als die Hälfte der Daten vorhanden waren.

Im ersten Schritt wurden die Zeitreihen in funktionale Daten umgewandelt. Dafür wurden Polynom-Splines gewählt, insbesondere Basis-Spline-Funktionen. Da die Anzahl der Knoten für die B-Splines die Anpassung der Funktion an die Daten bestimmt, wurden verschiedene Knotenmengen gewählt: jeweils eine mit 12 (vierteljährlich), 16 (alle zwei Monate), 36 (monatlich) und 72 (alle zwei Wochen) inneren Knoten. Diese Knoten sind äquidistant auf dem Gesamtintervall verteilt. Die Approximationen mit monatlichen Knoten ergeben glatte und übersichtliche Funktionen. Werden mehr Knoten für die Schätzung verwendet, ähneln die Funktionen den Liniendiagrammen und werden unübersichtlich. Bei wenigen Knoten (12 bzw. 16 Knoten) verliert man lokale Informationen für die Approximation.

Ausgehend von der Schätzung mit den B-Splines mit 36 inneren Knoten wurde untersucht, inwiefern man die steirischen Messstationen zu Gruppen mit ähnlichen Messwerten zusammenfassen kann. Dabei wurden die drei folgenden Clusteranalyseverfahren verwendet:

1. *k-means*-Verfahren, angewandt auf die geschätzten Splinekoeffizienten
2. *k-means*-Verfahren, angewandt auf die Beobachtungswerte
3. *Partitioning Around Medoids*-Verfahren, angewandt auf die geschätzten Splinekoeffizienten

Den Ergebnissen kann man folgende Beobachtungen entnehmen:

- Wenn nur wenige Objekte vorliegen, wie bei den Luftschadstoffen Ozon und Schwefeldioxid, dann bilden die drei Methoden dieselben Cluster.
- Das *k-means*-Verfahren, das die geschätzten Splinekoeffizienten der funktionalen Daten clustert, unterscheidet sich bei der Zuordnung der Messstationen zu den Clustern am stärksten von den beiden anderen Methoden.
- *Partitioning Around Medoids*-Verfahren eignen sich für viele Objekte, die sehr nahe beieinander liegen. Diese Methode geht von „kugelförmigen" Clustern aus. Für die steirischen Messstationen ergibt sich die Schwierigkeit, dass einerseits die Entfernungen zwischen den Stationen relativ groß sind und andererseits die Anzahl der verfügbaren Messstellen gering ist. Dies ist bei der Interpretation der Ergebnisse des *PAM*-Verfahrens zu berücksichtigen.

In Tabelle 5.1 sind für jeden Luftschadstoff und für jedes Verfahren die Anzahl der Messstationen pro Cluster aufgelistet, wobei zusätzlich die Methoden für unterschiedliche Klassenanzahl ($k = 2, \ldots, 4$) angewandt wurden. Mit Rot unterlegt sind die Cluster, die die Messstationen mit den höchsten Daten zusammenfassen, und mit Blau jene, die die niedrigsten Werte enthalten.

	k-means, Spline									*k-means, Rohdaten*									*PAM*								
	k = 2		*k = 3*			*k = 4*				*k = 2*		*k = 3*			*k = 4*				*k = 2*		*k = 3*			*k = 4*			
PM_{10}	4	20	4	16	4	4	11	4	5	12	12	4	8	12	4	9	4	7	4	20	3	9	12	3	8	6	7
O_3	1	9	1	6	3	1	6	2	1	1	9	1	6	3	1	3	3	3	1	9	3	9	12	1	6	2	1
SO_2	1	16	1	2	14	1	1	14	1	1	16	1	2	14	1	1	14	1	1	16	1	2	14	1	1	14	1
NO_2	3	21	3	20	1	1	10	1	12	3	21	1	10	13	1	10	1	12	12	12	1	11	12	1	11	1	11

Tab. 5.1. Anzahl der Objekte pro Cluster bei unterschiedlichen Clusteranzahl *k*

Bei den Luftschadstoffen Ozon und Schwefeldioxid gibt es bei jedem Verfahren und für jede Clusteranzahl (mindestens) eine Gruppe, die nur ein Element enthält. Das liegt zum einen daran, dass die Anzahl der Messstationen sehr gering ist und zum anderen an sehr auffallenden Messwerten (Masenberg bei Ozon und Straßengel bei Schwefeldioxid). Beim Luftschadstoff Stickstoffdioxid wird ab einer Aufteilung in 3 Cluster eine

Gruppe mit nur einer Messstation erzeugt. In dieser liegt entweder die Station mit den höchsten Werten (Graz Don Bosco) oder mit den niedrigsten Messwerten (Masenberg). Bei der Aufspaltung in 4 Cluster werden unabhängig von der Methode zwei Cluster mit jeweils einem Objekt gebildet.

Betrachtet man die Verteilung der Messstationen mit hohen und niedrigen Schadstoffkonzentrationen in der Steiermark, wird deutlich, dass die nördlichen und nord-östlichen Regionen weniger hohe Werte bei den Luftschadstoffen Feinstaub, Schwefeldioxid und Stickstoffdioxid aufweisen (siehe Abbildung 5.1, 5.3 und 5.4). Die Messstationen Liezen, Mürzzuschlag, Masenberg und Fürstenfeld werden dieser Gruppe bei allen drei Verfahren zugeordnet. Bei den Messstellen in der Mur-Mürz-Furche (Judenburg, Zeltweg, Knittelfeld, Leoben/Donawitz, Leoben/Göß, Niklasdorf, Bruck a. d. Mur und Kapfenberg) erkennt man Unterschiede bzgl. der Methoden. Das *PAM*-Verfahren ordnet die meisten Stationen der niedrigen Klasse zu, das *k-means*-Verfahren (Splinekoeffizienten) teilt diese eher der mittlere Gruppe zu (siehe Abbildung 5.1 und 5.4).

Bei Feinstaub finden sich die höchsten Messwerte in Graz (Graz Don Bosco und Graz Süd) und im Süden der Steiermark (Leibnitz). Diese Messstellen sind durch ihre kesselartige Lage gekennzeichnet. Zudem ist Graz Don Bosco ein wichtiger Verkehrsknotenpunkt. Bei Stickstoffdioxid unterscheidet sich die Zuordnung zur Gruppe mit den höchsten Werten stärker. Mit *PAM* ist nur die Messstelle Graz Don Bosco in diesem Cluster, nach dem *k-means*-Verfahren (Splinekoeffizienten) werden noch Graz Süd und Graz West dort zugeteilt. Bei Schwefeldioxid besteht das Cluster mit den höchsten Werten nur aus einem Messort, und zwar Straßengel. Die Messwerte dieser Messstation weichen derart stark von den übrigen ab, dass sie ein eigenes Cluster bildet.

Für den Luftschadstoff Ozon sind 10 Messstellen in der Untersuchung. Die Aufteilung in drei Cluster ist eindeutig. Hier erzeugt jedes Verfahren dieselbe Partition: Das Cluster mit den höchsten Werten enthält Masenberg (im Osten der Steiermark), im Cluster mit den niedrigsten Werten sind Liezen, Judenburg und Mürzzuschlag (im Norden der Steiermark) und die restlichen Stationen mit den mittleren Werten befinden sich im Süden bzw. Südwesten der Steiermark. Masenberg ist die höchstgelegene Messstation der Steiermark (1180 m Seehöhe), das heißt, dass die Vegetation und die Seehöhe einen Teil der Erklärung der hohen Messwerte ausmachen werden.

Bei den Clustern der mittleren Werte ist die Zuordnung für jeden Luftschadstoff unterschiedlich, vor allem wenn diese in 2 Gruppen aufgespalten werden (Clusterung mit

vier Clustern). Man kann das sehr gut bei Feinstaub und Stickstoffdioxid erkennen. Die Verfahren ordnen die Messstationen unterschiedlich zu Clustern zu.

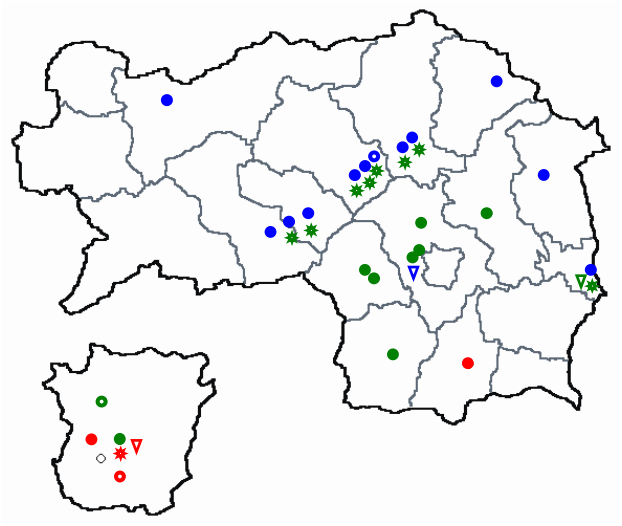


Abb. 5.1. PM_{10} Cluster der Steiermark

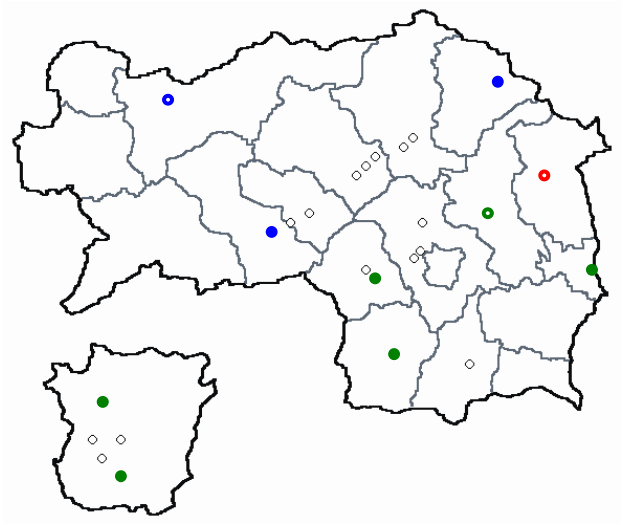


Abb. 5.2. O_3 Cluster der Steiermark

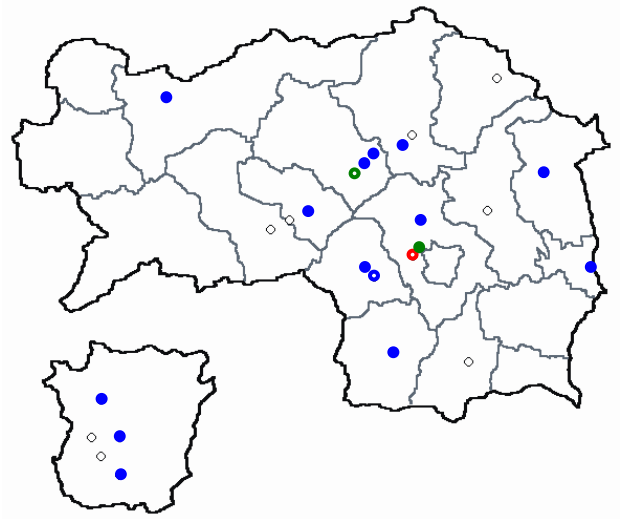


Abb. 5.3. SO_2 Cluster der Steiermark

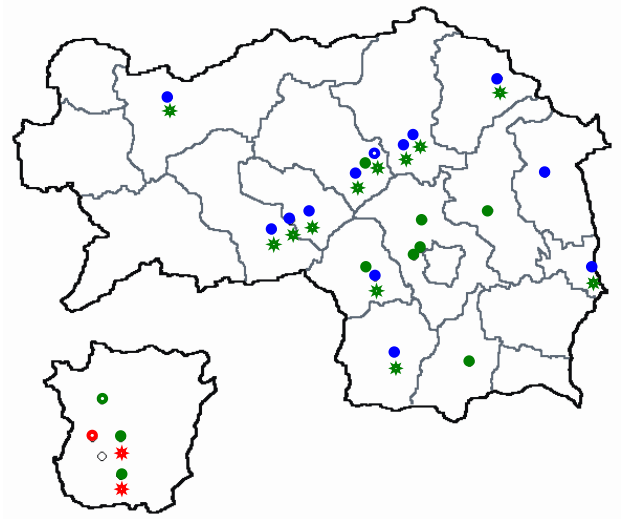


Abb. 5.4. NO_2 Cluster der Steiermark

Zusammenfassend kann man die Steiermark in einen nördlichen Bereich mit niedrigeren und in einen südlichen mit höheren Schadstoffkonzentrationen einteilen, wobei die höchsten Konzentrationen im Raum Graz auftreten. Vor allem die Grazer Messstationen, insbesondere die Messstelle am Verkehrsknotenpunkt Graz Don Bosco, zählen zu den Stationen mit hohen Werten. Liezen, Mürzzuschlag, Masenberg und Fürstenfeld befinden sich am Rand der Steiermark und zum Teil in höheren Regionen. Hier werden die niedrigsten Werte gemessen (bis auf die Ozonwerte bei Masenberg). Die übrigen Messstationen liegen zum Großteil im mittleren Cluster, jedoch werden die Stationen im Norden eher zu den niedrigen gezählt (Cluster 4) und die südlicheren zu den höheren (Cluster 2).

VI

LITERATURVERZEICHNIS

Bacher, J. (1996): Clusteranalyse. Anwendungsorientierte Einführung. München, Wien: Oldenbourg.

Fahrmeir, L, Kneib, T., Lang, S. (2007): Regression. Modelle, Methoden und Anwendungen. Berlin, Heidelberg, New York: Springer.

Ferraty, F., Vieu, P. (2006): Nonparametric Functional Data Analysis. Theory and Practice. New York: Springer.

Ignaccolo, R., Ghigo, S., Giovenali, E. (2008): Analysis of air quality monitoring networks by functional clustering. *Environmetrics 19*, 672 – 686.

Kaufman, L., Rousseeuw, P. J. (1990): Finding Groups in Data. An Introduction to Cluster Analysis. New York: John Wiley & Sons.

Öllinger, B. (2010): Funktionales Clustern von Transaktionsverläufen. Universität München. Institut für Statistik. Hochschulschrift.

Pruscha, H. (2006): Statistisches Methodenbuch. Verfahren, Fallstudien, Programmcodes. Berlin: Springer.

Ramsay, J. O., Hooker, G., Graves, S. (2009): Functional Data Analysis with R and MATLAB. Heidelberg: Springer.

Ramsay, J. O., Silverman, B. W. (1997): Functional Data Analysis. New York: Springer.

Rousseeuw, P. J. (1987): Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics 20*, 53 – 65.

Späth, H. (1977): Cluster-Analyse-Algorithmen zur Objektklassifizierung und Datenreduktion. 2. verb. Aufl. München: Oldenbourg.

[FA 17C, 2010/1] Antrag auf Fristerstreckung zur Einhaltung der PM10-Grenzwerte im Sanierungsgebiet Großraum Graz gemäß Artikel 22 Abs. 2 der Richtlinie 2008/50/EG. Bericht LU-01-2010. Amt der Steiermärkischen Landesregierung. Fachabteilung 17C. Februar 2010. [http://www.umwelt.steiermark.at/, am 4. 4. 2011]

[FA 17C, Juli 2010/2] Statuserhebung NO2 in Graz 2003 – 2009 gemäß §8 Immissionsschutzgesetz Luft BGBl. I Nr. 115/1997 i.d.g.F. Bericht: Lu-02-2010. Amt der Steiermärkischen Landesregierung. Fachabteilung 17C. Juli 2010. [http://www.umwelt.steiermark.at/, am 4. 4. 2011]

LUIS – Landes-Umwelt-Informationssystem. [http://www.umwelt.steiermark.at]

The comprehensive R Archive Network. R.2.14.2 for Windows [http://www.cran.r-project.org]

VII

ANHANG

A Tabellen

A.1 Tabellen zur Beschreibung der Messstationen

Bezirk	*Codierung*	*Stationen*	*Seehöhe*	*Anzahl*
Bruck	1 6	Bruck a. d. Mur Westend Kapfenberg	495 m 517 m	2
Deutschlandsberg	2	Deutschlandsberg AK	368 m	1
Feldbach		–		0
Fürstenfeld	3	Fürstenfeld	280 m	1
Graz	21 22 23 24 25	Graz Don Bosco Graz Mitte Gries Graz Nord Graz Süd Tiergartenweg Graz West	358 m 350 m 348 m 340 m 370 m	5
Graz Umgebung	5 16 17	Judendorf-Süd Peggau Straßengel-Kirche	375 m 410 m 454 m	3
Hartberg	13	Masenberg	1180 m	1
Judenburg	4 20	Judenburg Zeltweg-Hauptschule	715 m 675 m	2
Knittelfeld	7	Knittelfeld-Parkstraße	635 m	1
Leibnitz	9	Leibnitz	274 m	1
Liezen	12	Liezen	665 m	1
Leoben	10 11 15	Leoben- Donawitz Leoben-Göß Niklasdorf	555 m 554 m 510 m	3
Murau		–		0
Mürzzuschlag	14	Mürzzuschlag-Roseggerpark	679 m	1
Radkersburg		–		0
Voitsberg	8 18	Köflach Voitsberg	445 m 390 m	2
Weiz	19	Weiz	479 m	1

Tab. A.1.1. Messstationen, nach Bezirken aufgelistet

A.2 Tabellen zu Clustern mit unterschiedlichen k

Die folgenden Tabellen zeigen die Zuordnung der Messstationen zu den Clustern bei Verwendung unterschiedlicher Klassenanzahl bzw. Methode. Um einen Vergleich zu ermöglichen, wurden die Cluster umnummeriert und farblich markiert. Cluster 1 enthält die Messstationen mit den höchsten Werten und Cluster 3 jene mit den niedrigsten Werten. In Cluster 2 bzw. Cluster 4 befinden sich die Messstationen mit den mittleren Werten (zweithöchsten bzw. zweitniedrigsten Werte).

A.2.1 Feinstaub PM_{10} [µg/m³]

k-means, Spline			*k-means, Rohdaten*			*PAM*		
k = 2	*k = 3*	*k = 4*	*k = 2*	*k = 3*	*k = 4*	*k = 2*	*k = 3*	*k = 4*
4 | 20	4 | 16 | 4	4 | 11 | 4 | 5	12 | 12	4 | 8 | 12	4 | 9 | 4 | 7	4 | 20	3 | 9 | 12	3 | 8 | 6 | 7

Tab. A.2.1 PM_{10}: Anzahl der Objekte pro Cluster für verschiedene Partitionen

	k-means, Spline			*k-means, Rohdaten*			*PAM*		
	k = 2	*k = 3*	*k = 4*	*k = 2*	*k = 3*	*k = 4*	*k = 2*	*k = 3*	*k = 4*
Bruck	2	2	4	2	3	4	2	3	4
Deutschlandsberg	2	2	2	1	2	2	2	2	2
Fürstenfeld	2	2	2	1	2	2	2	3	4
Judenburg	2	3	3	2	3	3	2	3	3
Judendorf	2	2	2	1	2	2	2	2	2
Kapfenberg	2	2	4	2	3	4	2	3	4
Knittelfeld	2	2	2	2	3	4	2	3	3
Köflach	2	2	2	1	2	2	2	2	4
Leibnitz	1	1	1	1	1	1	1	1	1
Leoben/Donawitz	2	2	4	2	3	4	2	3	4
Leoben/Göß	2	2	4	2	3	4	2	3	4
Liezen	2	3	3	2	3	3	2	3	3
Masenberg	2	3	3	2	3	3	2	3	3
Mürzzuschlag	2	3	3	2	3	3	2	3	3
Niklasdorf	2	2	4	2	3	4	2	3	4
Peggau	2	2	2	1	2	2	2	2	2
Straßengel	2	2	2	2	3	2	2	2	2
Voitsberg	2	2	2	1	2	2	2	2	2
Weiz	2	2	2	1	2	2	2	2	2
Zeltweg	2	2	2	2	3	4	2	3	3
Graz Don Bosco	1	1	1	1	1	1	1	1	1
Graz Nord	2	2	2	1	2	2	2	2	2
Graz Süd	1	1	1	1	1	1	1	1	1
Graz West	1	1	1	1	1	1	1	2	2

Tab. A.2.2. PM_{10}: Zuordnung der Stationen zu den einzelnen Klassen

A.2.2 Ozon O_3 [µg/m³]

k-means, Spline			*k-means, Rohdaten*			*PAM*		
k = 2	*k = 3*	*k = 4*	*k = 2*	*k = 3*	*k = 4*	*k = 2*	*k = 3*	*k = 4*
1 | 9	1 | 6 | 3	1 | 6 | 2 | 1	1 | 9	1 | 6 | 3	1 | 3 | 3 | 3	1 | 9	3 | 9 | 12	1 | 6 | 2 | 1

Tab. A.2.3 O_3: Anzahl der Objekte pro Cluster für verschiedene Partitionen

	k-means, Spline			*k-means, Rohdaten*			*PAM*		
	k = 2	*k = 3*	*k = 4*	*k = 2*	*k = 3*	*k = 4*	*k = 2*	*k = 3*	*k = 4*
Deutschlandsberg	2	2	2	2	2	2	2	2	2
Fürstenfeld	2	2	2	2	2	2	2	2	2
Judenburg	2	3	4	2	3	3	2	3	4
Liezen	2	3	3	2	3	3	2	3	3
Masenberg	1	1	1	1	1	1	1	1	1
Mürzzuschlag	2	3	3	2	3	3	2	3	3
Voitsberg	2	2	2	2	2	4	2	2	2
Weiz	2	2	2	2	2	2	2	2	2
Graz Nord	2	2	2	2	2	4	2	2	2
Graz Süd	2	2	2	2	2	4	2	2	2

Tab. A.2.4. O_3: Zuordnung der Stationen zu den einzelnen Klassen

A.2.3 Schwefeldioxid SO_2 [µg/m³]

k-means, Spline			*k-means, Rohdaten*			*PAM*		
k = 2	*k = 3*	*k = 4*	*k = 2*	*k = 3*	*k = 4*	*k = 2*	*k = 3*	*k = 4*
1 | 16	1 | 2 | 14	1 | 1 | 14 | 1	1 | 16	1 | 2 | 14	1 | 1 | 14 | 1	1 | 16	1 | 2 | 14	1 | 1 | 14 | 1

Tab. A.2.5 SO_2: Anzahl der Objekte pro Cluster für verschiedene Partitionen

	k-means, Spline			*k-means, Rohdaten*			*PAM*		
	k = 2	*k = 3*	*k = 4*	*k = 2*	*k = 3*	*k = 4*	*k = 2*	*k = 3*	*k = 4*
Bruck	2	3	3	2	3	3	2	3	3
Deutschlandsberg	2	3	3	2	3	3	2	3	3
Fürstenfeld	2	3	3	2	3	3	2	3	3
Judendorf	2	2	4	2	2	4	2	2	4
Knittelfeld	2	3	3	2	3	3	2	3	3
Köflach	2	3	3	2	3	3	2	3	3
Leoben/Donawitz	2	2	2	2	2	2	2	2	2
Leoben/Göß	2	3	3	2	3	3	2	3	3
Liezen	2	3	3	2	3	3	2	3	3
Masenberg	2	3	3	2	3	3	2	3	3
Niklasdorf	2	3	3	2	3	3	2	3	3
Peggau	2	3	3	2	3	3	2	3	3
Straßengel	1	1	1	1	1	1	1	1	1
Voitsberg	2	3	3	2	3	3	2	3	3
Graz Nord	2	3	3	2	3	3	2	3	3
Graz Süd	2	3	3	2	3	3	2	3	3
Graz West	2	3	3	2	3	3	2	3	3

Tab. A.2.6. SO_2: Zuordnung der Stationen zu den einzelnen Klassen

A.2.4 Stickstoffdioxid NO_2 [µg/m³]

k-means, Spline			*k-means, Rohdaten*			*PAM*		
k = 2	*k = 3*	*k = 4*	*k = 2*	*k = 3*	*k = 4*	*k = 2*	*k = 3*	*k = 4*
3 | 21	3 | 20 | 1	1 | 10 | 1 | 12	3 | 21	1 | 10 | 13	1 | 10 | 1 | 12	12 | 12	1 | 11 | 12	1 | 11 | 1 | 11

Tab. A.2.7 NO_2: Anzahl der Objekte pro Cluster für verschiedene Partitionen

	k-means, Spline			*k-means, Rohdaten*			*PAM*		
	k = 2	*k = 3*	*k = 4*	*k = 2*	*k = 3*	*k = 4*	*k = 2*	*k = 3*	*k = 4*
Bruck	2	2	4	2	3	4	2	3	4
Deutschlandsberg	2	2	4	2	3	4	2	3	4
Fürstenfeld	2	2	4	2	3	4	2	3	4
Judenburg	2	2	4	2	3	4	2	3	4
Judendorf	2	2	2	2	2	2	1	2	2
Kapfenberg	2	2	4	2	3	4	2	3	4
Knittelfeld	2	2	4	2	3	4	2	3	4
Köflach	2	2	2	2	2	2	1	2	2
Leibnitz	2	2	2	2	2	2	1	2	2
Leoben/Donawitz	2	2	4	2	3	4	1	2	2
Leoben/Göß	2	2	2	2	2	2	1	2	2
Liezen	2	2	4	2	3	4	2	3	4
Masenberg	2	3	3	2	3	3	2	3	3
Mürzzuschlag	2	2	4	2	3	4	2	3	4
Niklasdorf	2	2	4	2	3	4	2	3	4
Peggau	2	2	2	2	2	2	1	2	2
Straßengel	2	2	2	2	2	2	1	2	2
Voitsberg	2	2	4	2	3	4	2	3	4
Weiz	2	2	2	2	2	2	1	2	2
Zeltweg	2	2	4	2	3	4	2	3	4
Graz Don Bosco	1	1	1	1	1	1	1	1	1
Graz Nord	2	2	2	2	2	2	1	2	2
Graz Süd	1	1	2	1	2	2	1	2	2
Graz West	1	1	2	1	2	2	1	2	2

Tab. A.2.8. NO_2: Zuordnung der Stationen zu den einzelnen Klassen

A.3 Silhouette-Breiten des *PAM*-Verfahrens

A.3.1 Feinstaub PM_{10} [µg/m³]

	Cluster	*Nachbar*	*s(i)*
Liezen	1	2	0.556
Niklasdorf	1	2	0.551
Zeltweg	1	2	0.536
Kapfenberg	1	2	0.532
Leoben/Göß	1	2	0.531
Mürzzuschlag	1	2	0.528
Judenburg	1	2	0.501
Bruck	1	2	0.484
Knittelfeld	1	2	0.458
Straßengel	1	2	0.452
Leoben/Donawitz	1	2	0.436
Masenberg	1	2	0.413
Köflach	1	2	0.374
Peggau	1	2	0.344
Fürstenfeld	1	2	0.326
Voitsberg	1	2	0.306
Judendorf	1	2	0.280
Weiz	1	2	0.251
Graz Nord	1	2	0.193
Deutschlandsberg	1	2	0.180
Graz Süd	2	1	0.595
Graz Don Bosco	2	1	0.570
Graz West	2	1	0.376
Leibnitz	2	1	0.295

Tab. A.3.1. PM_{10}: Silhouette-Breiten bei einer Clusterung mit 2 Repräsentanten

	Cluster	*Nachbar*	*s(i)*
Liezen	1	2	0.37
Mürzzuschlag	1	2	0.32
Niklasdorf	1	2	0.31
Leoben/Göß	1	2	0.29
Masenberg	1	2	0.26
Judenburg	1	2	0.23
Bruck	1	2	0.21
Kapfenberg	1	2	0.20
Zeltweg	1	2	0.14
Leoben/Donawitz	1	2	0.14
Knittelfeld	1	2	0.01
Fürstenfeld	1	2	-0.15
Graz Nord	2	1	0.41
Weiz	2	1	0.33
Judendorf	2	1	0.30
Deutschlandsberg	2	1	0.30
Voitsberg	2	1	0.25
Peggau	2	1	0.21
Köflach	2	1	0.19
Straßengel	2	1	0.05
Graz West	2	3	-0.18
Graz Süd	3	2	0.52
Graz Don Bosco	3	2	0.47
Leibnitz	3	2	0.06

Tab. A.3.2. PM_{10}: Silhouette-Breiten bei einer Clusterung mit 3 Repräsentanten

A.3.2 Ozon O_3 [µg/m³]

	Cluster	*Nachbar*	*s(i)*
Voitsberg	1	2	0.780
Weiz	1	2	0.766
Graz Nord	1	2	0.764
Graz Süd	1	2	0.760
Fürstenfeld	1	2	0.751
Deutschlandsberg	1	2	0.728
Mürzzuschlag	1	2	0.728
Liezen	1	2	0.703
Judenburg	1	2	0.686
Masenberg	2	1	0.000

Tab. A.3.3. O_3: Silhouette-Breiten bei einer Clusterung mit 2 Repräsentanten

	Cluster	*Nachbar*	*s(i)*
Graz Nord	1	2	0.415
Graz Süd	1	2	0.341
Weiz	1	2	0.318
Voitsberg	1	2	0.292
Deutschlandsberg	1	2	0.281
Fürstenfeld	1	2	0.264
Liezen	2	1	0.237
Mürzzuschlag	2	1	0.039
Judenburg	2	1	0.037
Masenberg	3	2	0.000

Tab. A.3.4. O_3: Silhouette-Breiten bei einer Clusterung mit 3 Repräsentanten

A.3.3 Schwefeldioxid SO_2 [µg/m³]

	Cluster	*Nachbar*	*s(i)*
Deutschlandsberg	1	2	0.853
Voitsberg	1	2	0.852
Peggau	1	2	0.848
Knittelfeld	1	2	0.846
Liezen	1	2	0.842
Köflach	1	2	0.841
Niklasdorf	1	2	0.836
Graz Nord	1	2	0.834
Leoben/Göß	1	2	0.834
Fürstenfeld	1	2	0.825
Bruck	1	2	0.822
Masenberg	1	2	0.820
Graz Süd	1	2	0.796
Graz West	1	2	0.792
Leoben/Donawitz	1	2	0.589
Judendorf	1	2	0.589
Straßengel	2	1	0.000

Tab. A.3.5. SO_2: Silhouette-Breiten bei einer Clusterung mit 2 Repräsentanten

	Cluster	*Nachbar*	*s(i)*
Voitsberg	1	2	0.647
Deutschlandsberg	1	2	0.629
Peggau	1	2	0.628
Knittelfeld	1	2	0.618
Köflach	1	2	0.601
Leoben/Göß	1	2	0.577
Liezen	1	2	0.576
Graz Nord	1	2	0.563
Niklasdorf	1	2	0.553
Masenberg	1	2	0.550
Fürstenfeld	1	2	0.515
Bruck	1	2	0.487
Graz Süd	1	2	0.389
Graz West	1	2	0.325
Leoben/Donawitz	2	1	0.084
Judendorf	2	1	0.010
Straßengel	3	2	0.000

Tab. A.3.6. SO_2: Silhouette-Breiten bei einer Clusterung mit 3 Repräsentanten

A.3.4 Stickstoffdioxid NO_2 [µg/m³]

	Cluster	*Nachbar*	*s(i)*
Niklasdorf	1	2	0.407
Deutschlandsberg	1	2	0.396
Judenburg	1	2	0.396
Bruck	1	2	0.295
Voitsberg	1	2	0.286
Knittelfeld	1	2	0.255
Masenberg	1	2	0.255
Zeltweg	1	2	0.224
Liezen	1	2	0.215
Kapfenberg	1	2	0.193
Mürzzuschlag	1	2	0.065
Fürstenfeld	1	2	–0.011
Graz West	2	1	0.447
Graz Nord	2	1	0.405
Straßengel	2	1	0.354
Leoben/Göß	2	1	0.331
Weiz	2	1	0.324
Graz Süd	2	3	0.320
Judendorf	2	1	0.263
Leibnitz	2	1	0.246
Peggau	2	1	0.218
Köflach	2	1	0.181
Leoben/Donawitz	2	1	0.017
Graz Don Bosco	3	2	0.000

Tab. A.3.7. NO_2: Silhouette-Breiten bei einer Clusterung mit 3 Repräsentanten

	Cluster	*Nachbar*	*s(i)*
Niklasdorf	1	2	0.511
Deutschlandsberg	1	2	0.471
Judenburg	1	2	0.457
Bruck	1	2	0.434
Voitsberg	1	2	0.404
Knittelfeld	1	2	0.377
Zeltweg	1	2	0.353
Liezen	1	2	0.327
Kapfenberg	1	2	0.294
Mürzzuschlag	1	2	0.220
Fürstenfeld	1	2	0.095
Graz West	2	1	0.395
Graz Nord	2	1	0.323
Graz Süd	2	4	0.321
Straßengel	2	1	0.280
Leoben/Göß	2	1	0.251
Weiz	2	1	0.235
Judendorf	2	1	0.150
Leibnitz	2	1	0.148
Peggau	2	1	0.099
Köflach	2	1	0.037
Leoben/Donawitz	2	1	–0.094
Masenberg	3	1	0.000
Graz Don Bosco	4	2	0.0000

Tab. A.3.8. NO_2: Silhouette-Breiten bei einer Clusterung mit 4 Repräsentanten

A.4 Randindices

A.4.1 Feinstaub PM_{10} [µg/m³]

	3 Cluster		*4 Cluster*	
	k-means (Rohdaten)	*PAM*	*k-means (Rohdaten)*	*PAM*
k-means (Spline)	1	0.33	0.22	0.76
k-means (Rohdaten)		0.33		0.17

Tab. A.4.1. PM_{10}: Randindex für den Vergleich der drei Cluster-Methoden

	k-means (Spline)		*PAM*	
	18 Knoten	*72 Knoten*	*18 Knoten*	*72 Knoten*
36 Knoten	1	0.33	0.22	0.76
18 Knoten		0.33		0.17

Tab. A.4.2. PM_{10}: Randindices bei einer Partition von drei Cluster mit unterschiedlicher Knotenanzahl

A.4.2 Ozon O_3 [µg/m³]

	3 Cluster		*4 Cluster*	
	k-means (Rohdaten)	*PAM*	*k-means (Rohdaten)*	*PAM*
k-means (Spline)	1	1	0.41	1
k-means (Rohdaten)		1		0.41

Tab. A.4.3. O_3: Randindex für den Vergleich der drei Cluster-Methoden

	k-means		*PAM*	
	18 Knoten	72 Knoten	18 Knoten	72 Knoten
36 Knoten	1	1	0.34	1
18 Knoten		1		0.34

Tab. A.4.4. O_3: Randindices bei einer Partition von drei Cluster mit unterschiedlicher Knotenanzahl

A.4.3 Schwefeldioxid SO_2 [µg/m³]

	3 Cluster		*4 Cluster*	
	k-means (Rohdaten)	*PAM*	*k-means (Rohdaten)*	*PAM*
k-means (Spline)	1	1	1	1
k-means (Rohdaten)		1		1

Tab. A.4.5. SO_2: Randindex für den Vergleich der drei Cluster-Methoden

	k-means		*PAM*	
	18 Knoten	*72 Knoten*	*18 Knoten*	*72 Knoten*
36 Knoten	1	1	0.73	0.73
18 Knoten		1		1

Tab. A.4.6. SO_2: Randindices bei einer Partition von drei Cluster mit unterschiedlicher Knotenanzahl

A.4.4 Stickstoffdioxid NO_2 [µg/m³]

	3 Cluster		*4 Cluster*	
	k-means (Rohdaten)	*PAM*	*k-means (Rohdaten)*	*PAM*
k-means (Spline)	0.12	0.1	1	0.84
k-means (Rohdaten)		0.84		0.84

Tab. A.4.7. NO_2: Randindex für den Vergleich der drei Cluster-Methoden

	k-means		*PAM*	
	18 Knoten	72 Knoten	18 Knoten	72 Knoten
36 Knoten	1	0.12	0.69	0.69
18 Knoten		0.12		1

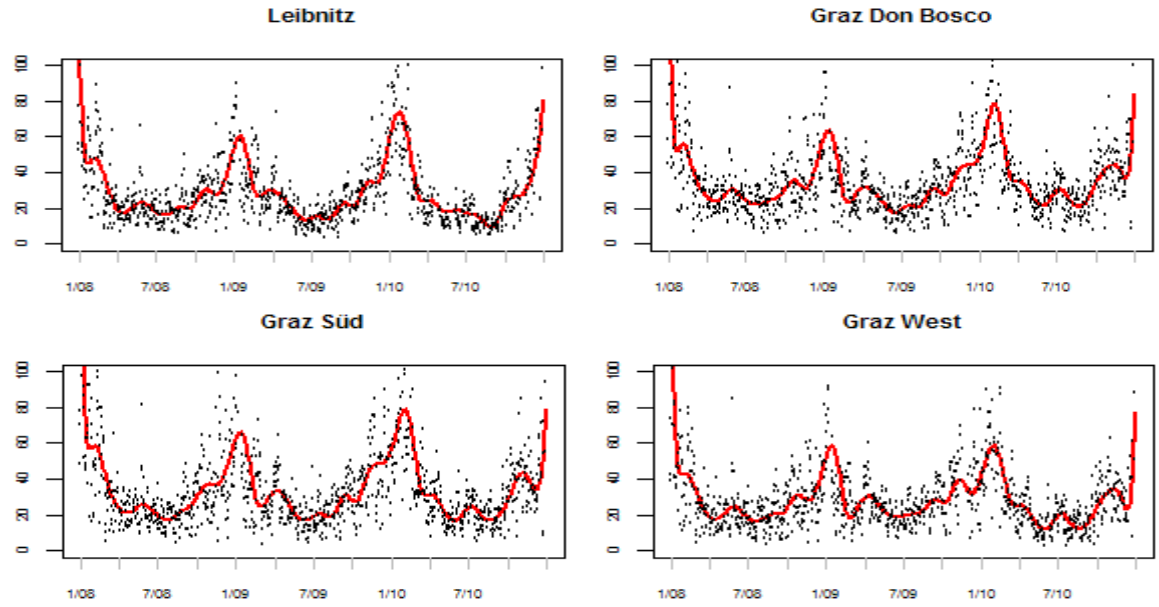
Tab. 4.3.8. NO_2: Randindices bei einer Partition von drei Cluster mit unterschiedlicher Knotenanzahl

B Graphiken

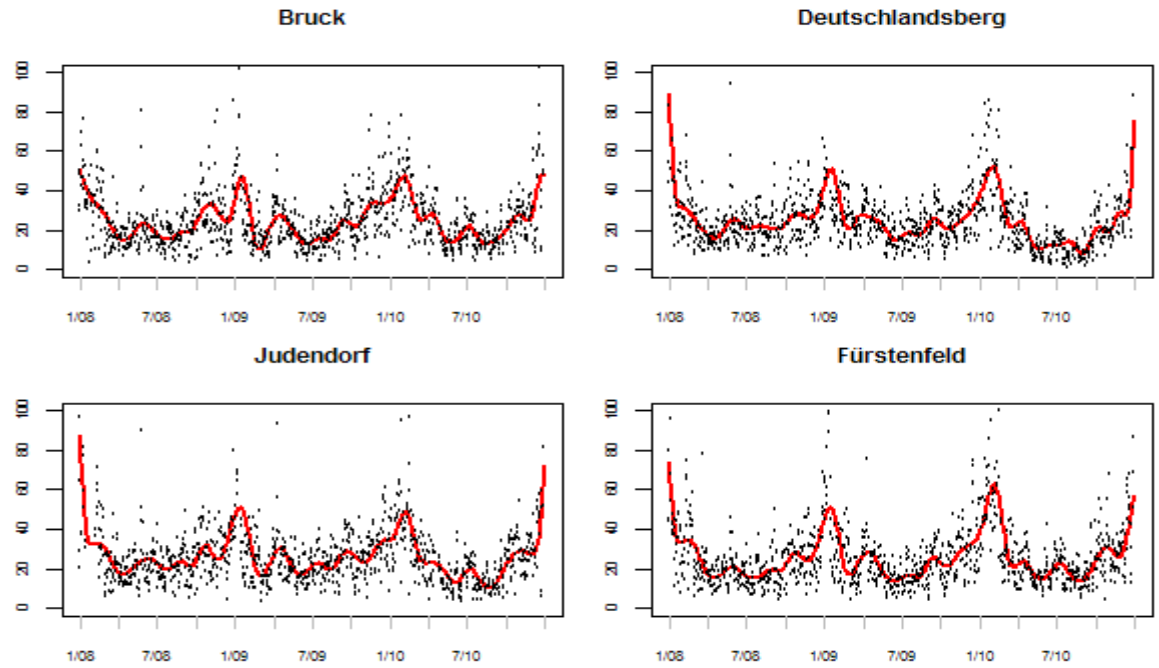
B.1 Approximationen der Funktionen

B.1.1 Feinstaub PM_{10} [µg/m³]

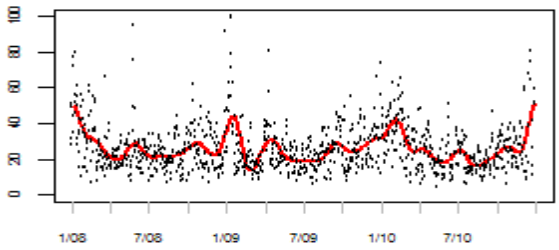
Cluster 1 (nach *k-means* der Splinekoeffizienten)



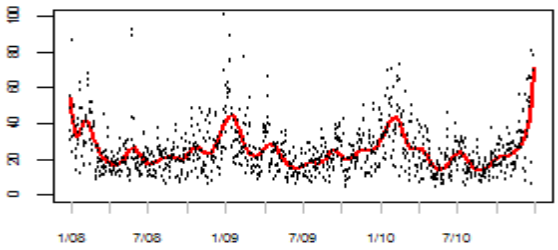
Cluster 2



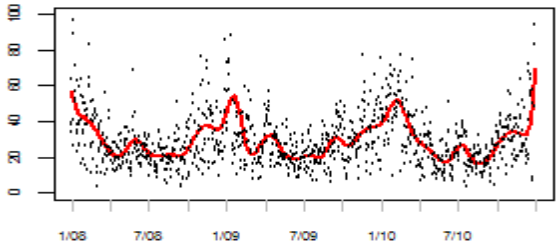
Kapfenberg



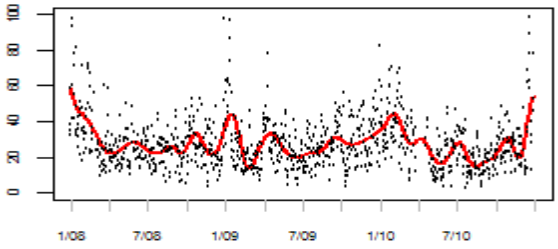
Knittelfeld



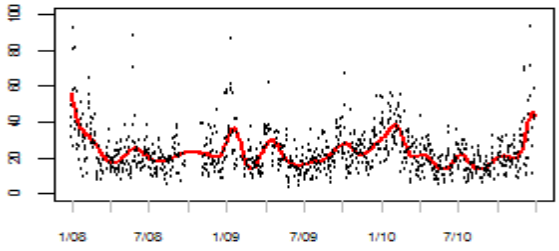
Köflach



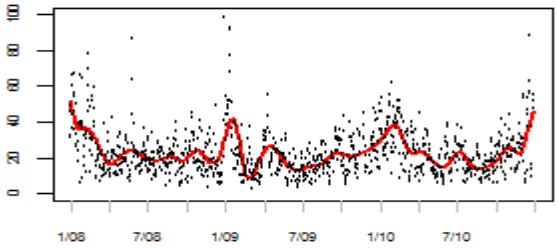
Leoben/Donawitz



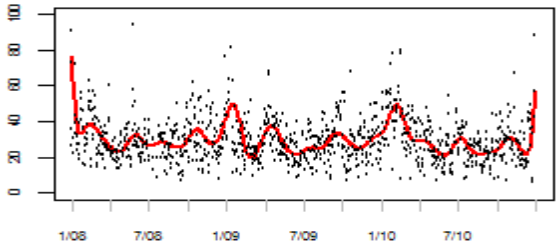
Leoben/Göß



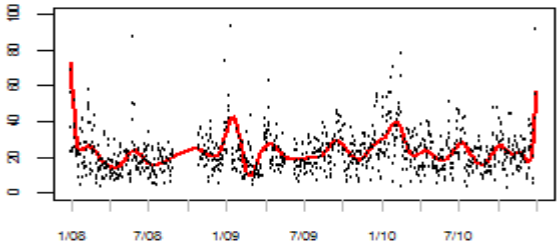
Niklasdorf



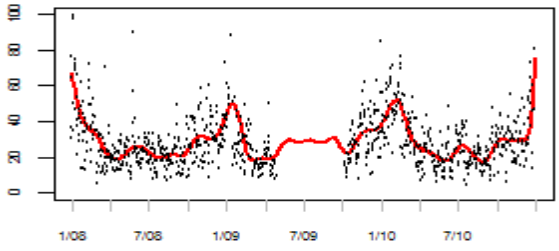
Peggau



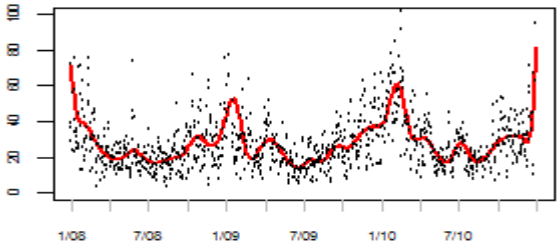
Straßengel



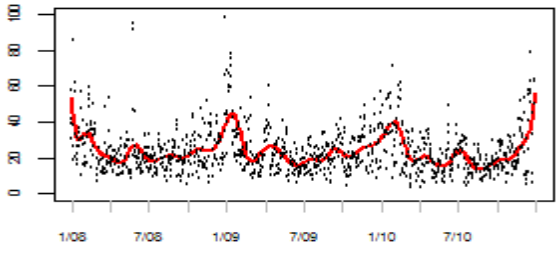
Voitsberg



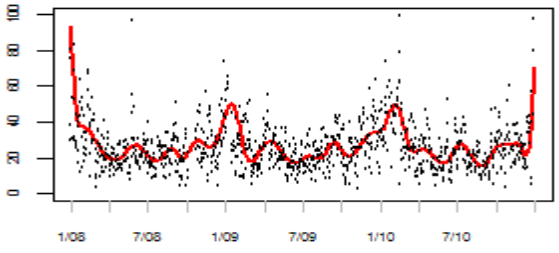
Weiz



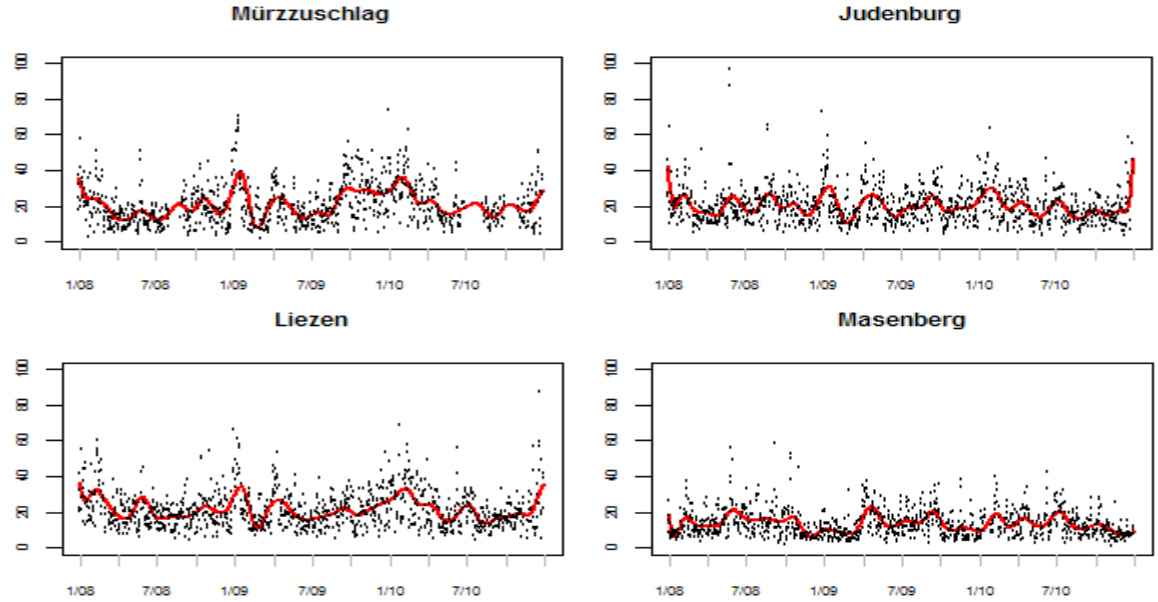
Zeltweg



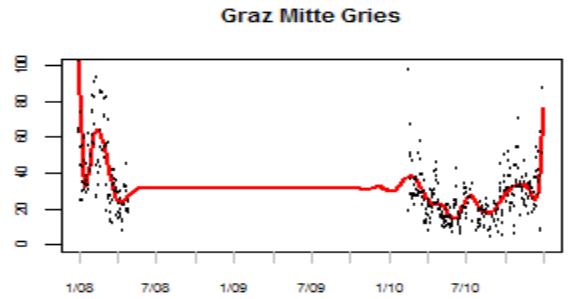
Graz Nord



Cluster 3

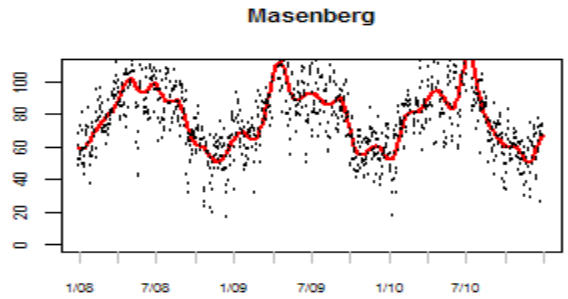


Approximation der Funktion für Graz Mitte Gries (diese Messstation ist nicht Teil der Clusterung)

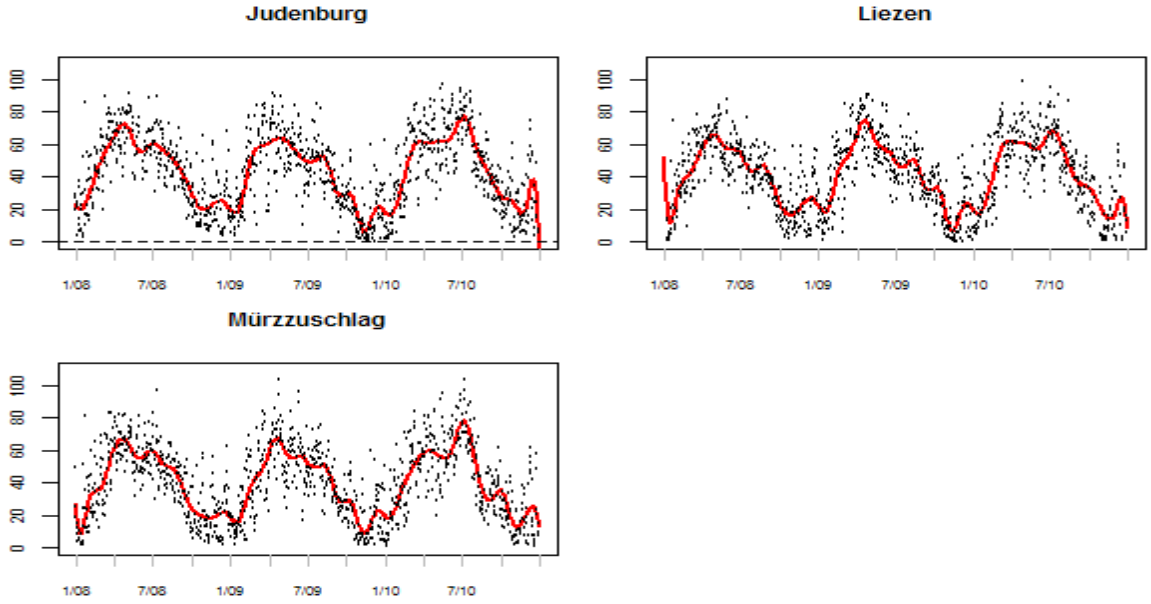


B.1.2 Ozon O_3 [µg/m³]

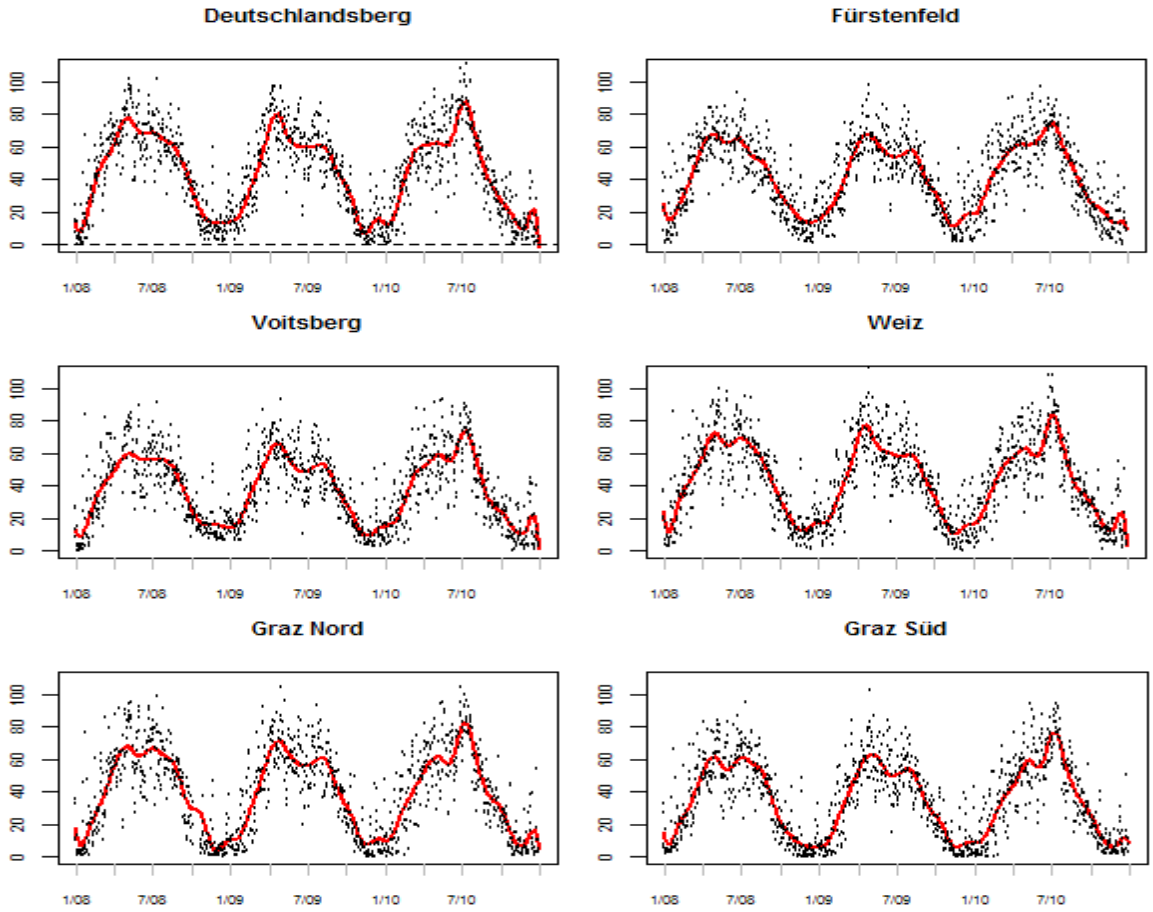
Cluster 1



Cluster 2

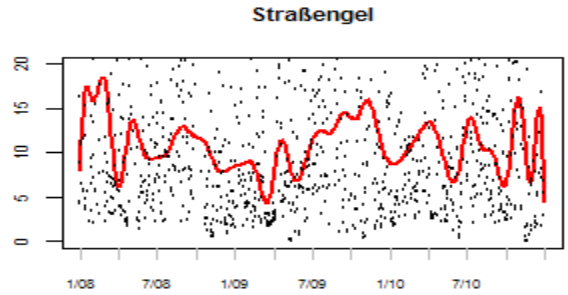


Cluster 3

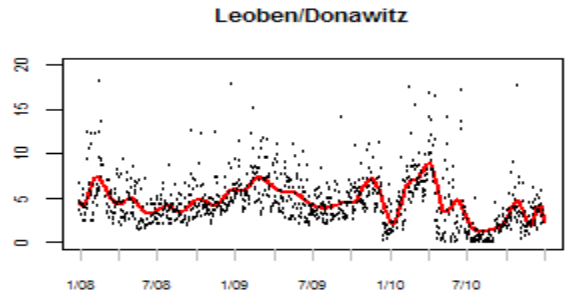


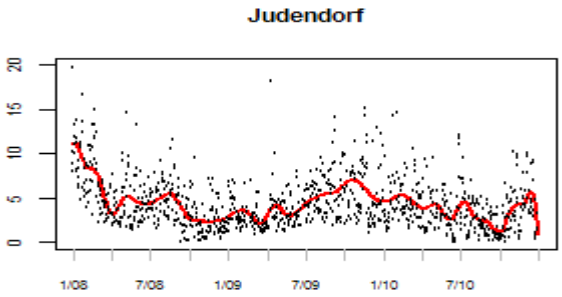
B.1.3 Schwefeldioxid SO_2 [µg/m³]

Cluster 1

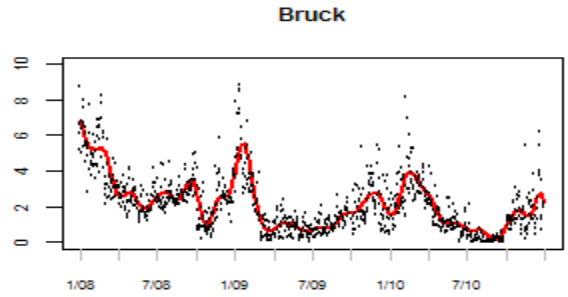


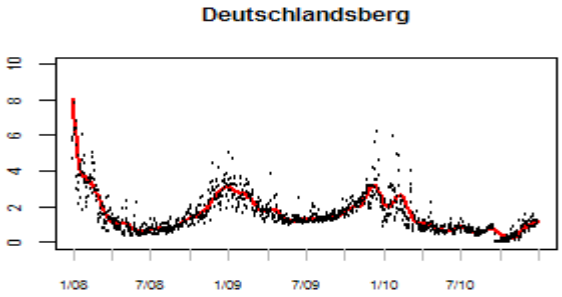
Cluster 2

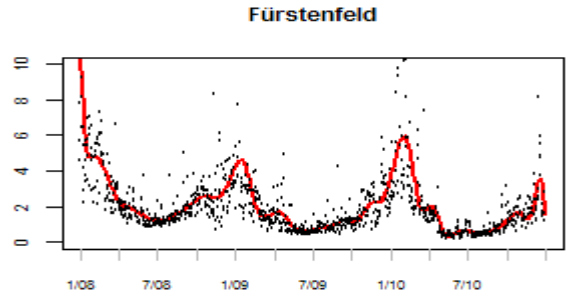


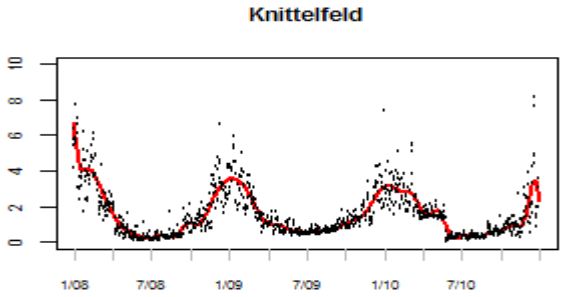


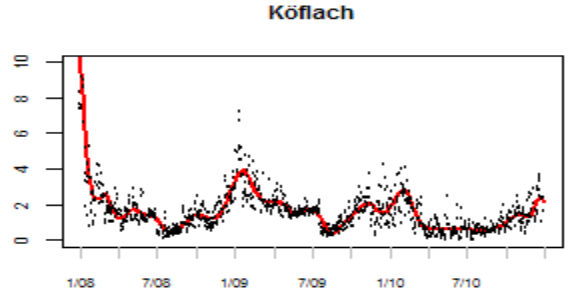
Cluster 3

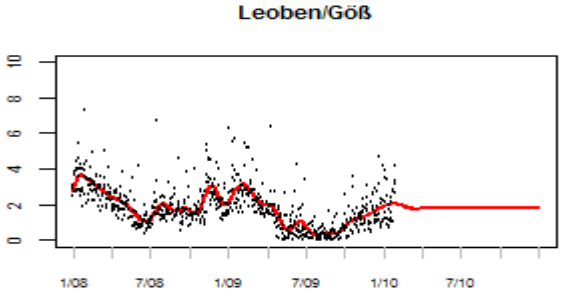


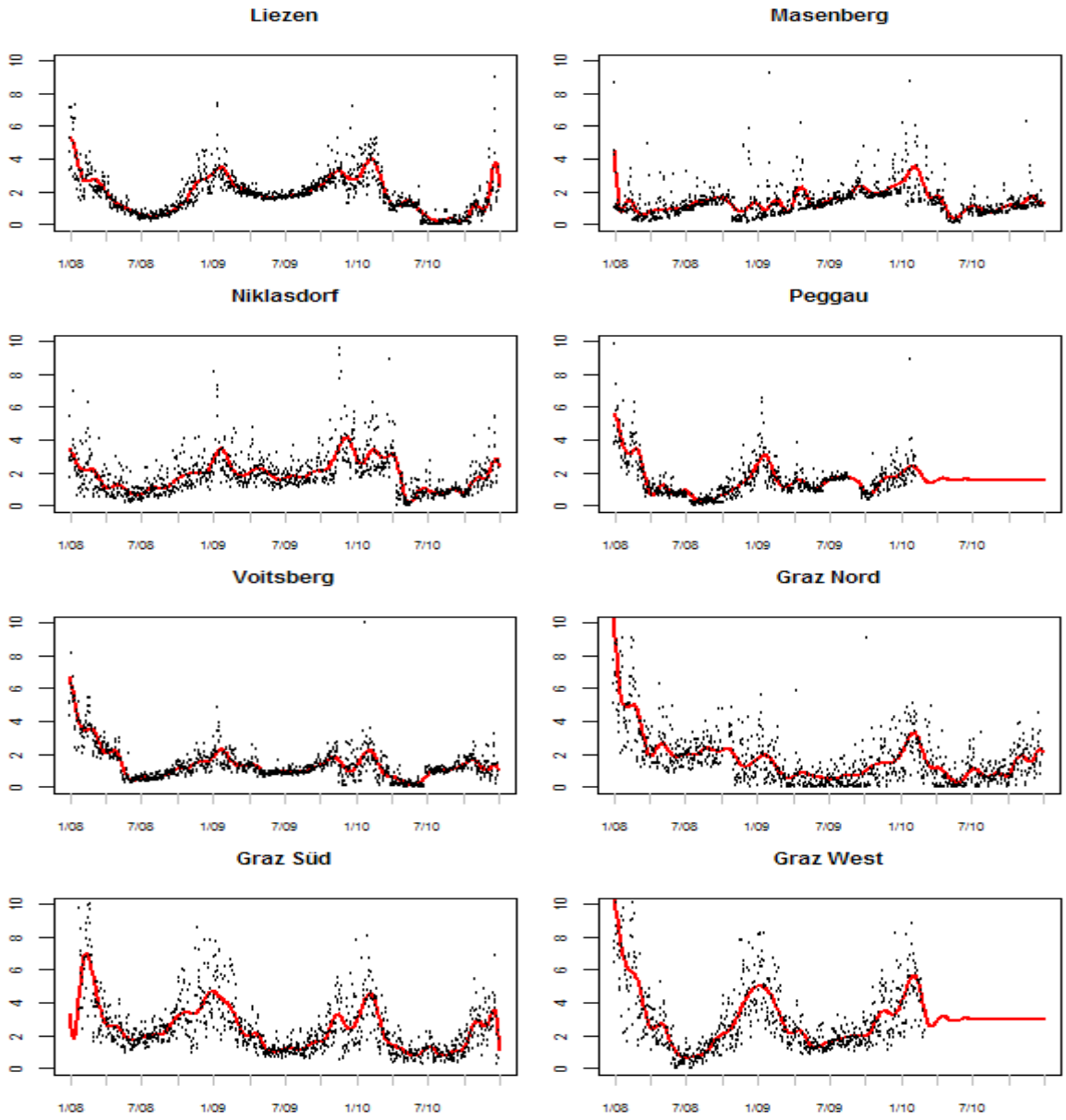




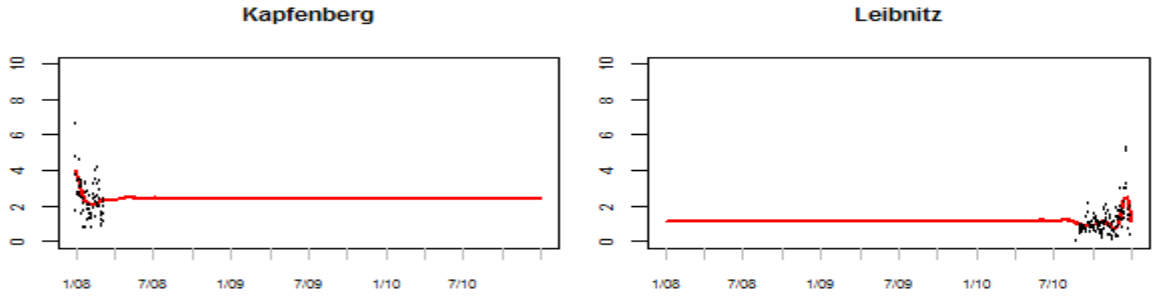






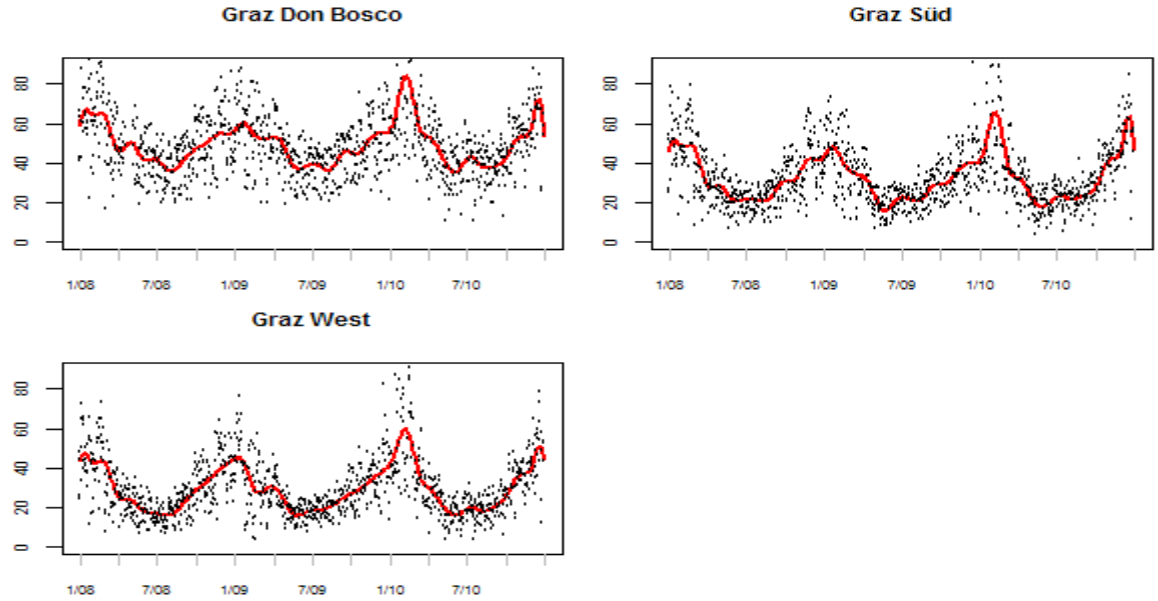


Approximation der Funktionen für Kapfenberg und Leibnitz (diese Messstationen sind nicht Teil der Clusterung)

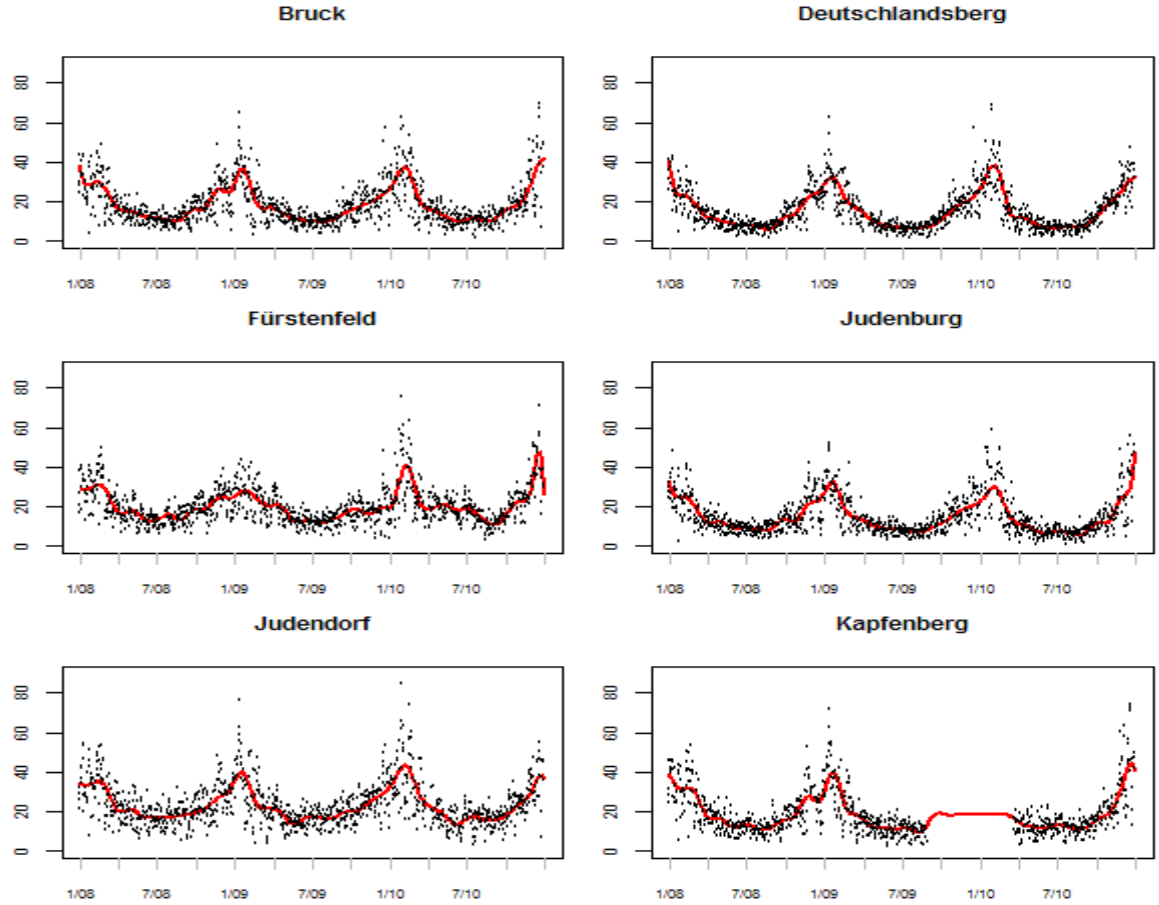


B.1.4 Stickstoffdioxid NO_2 [µg/m³]

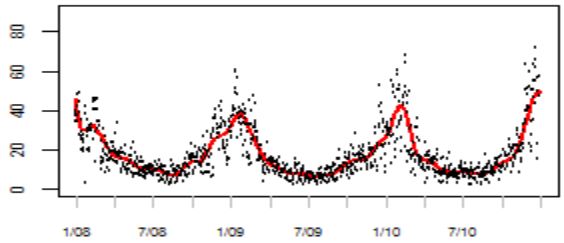
Cluster 1



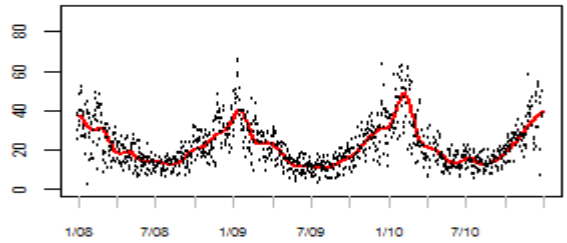
Cluster 2



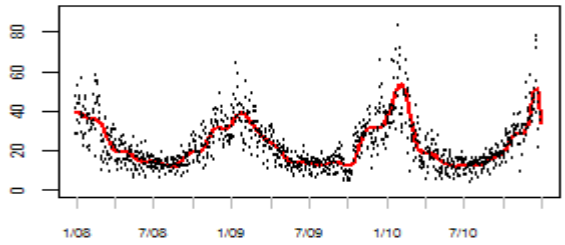
Knittelfeld



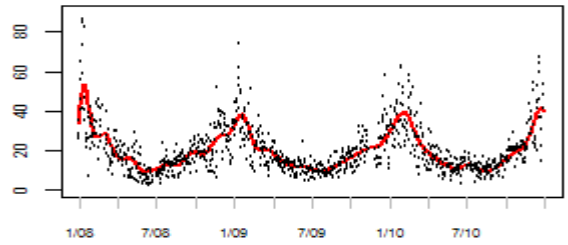
Köflach



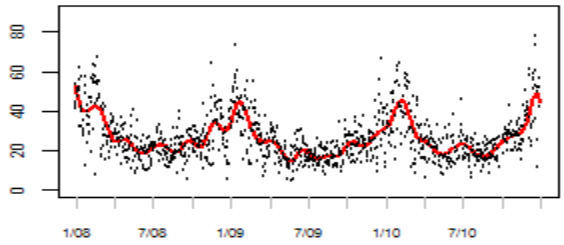
Leibnitz



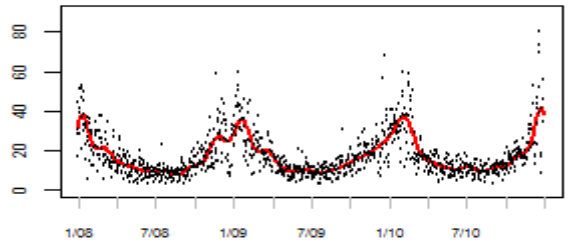
Leoben/Donawitz



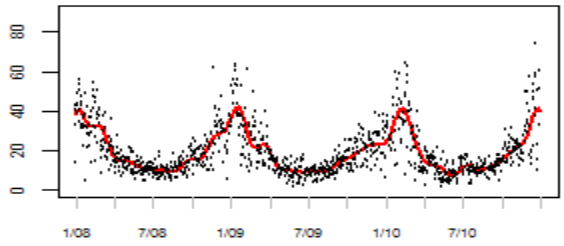
Leoben/Göß



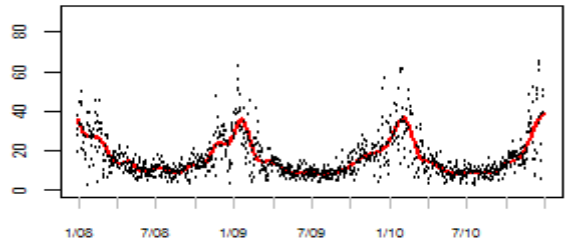
Liezen



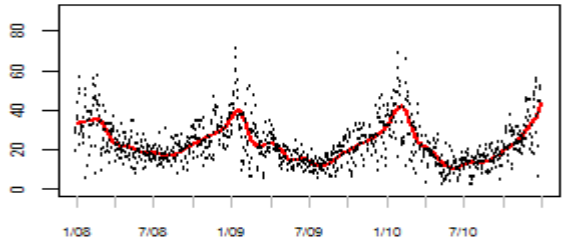
Mürzzuschlag



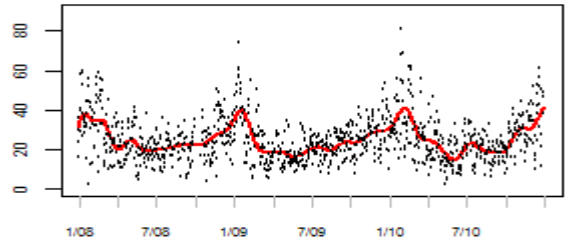
Niklasdorf



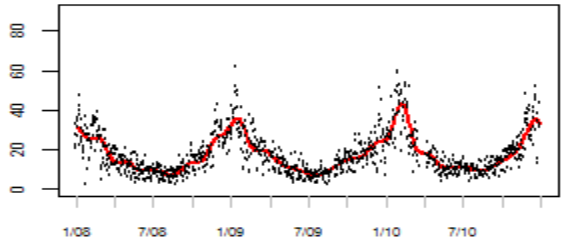
Peggau



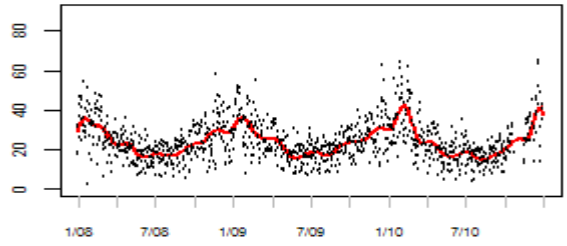
Straßengel

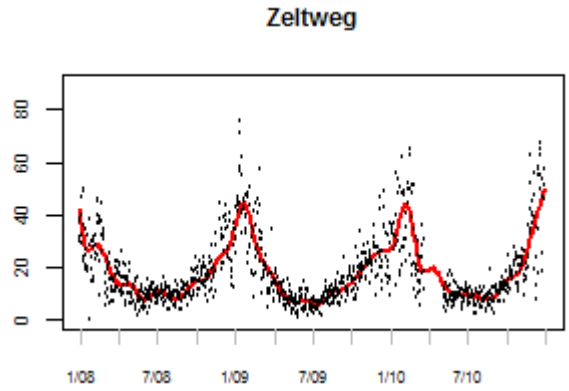


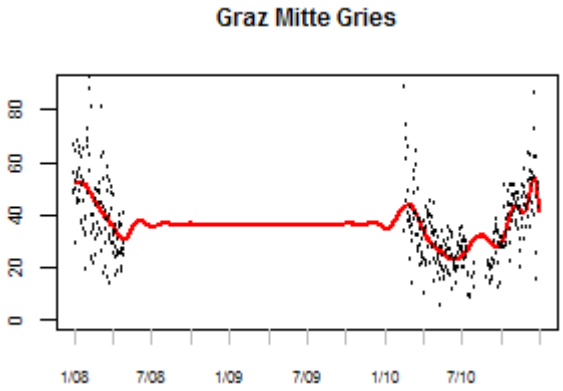
Voitsberg

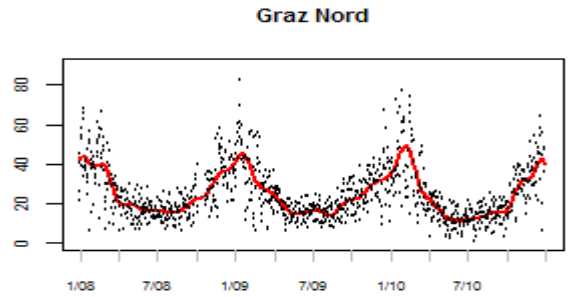


Weiz

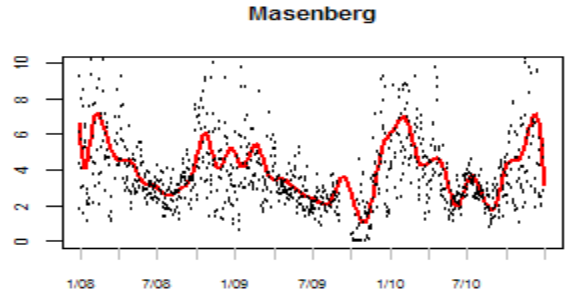




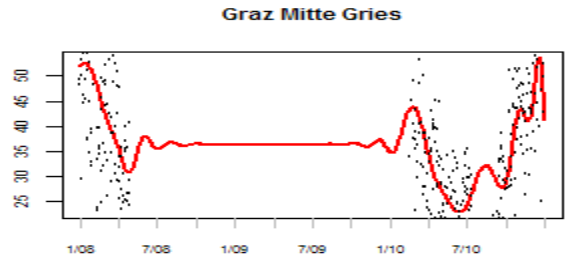




Cluster 3



Approximation der Funktion für Graz Mitte Gries (diese Messstation ist Teil der Clusterung)



B.2 Screeplots und Silhouette-Breiten

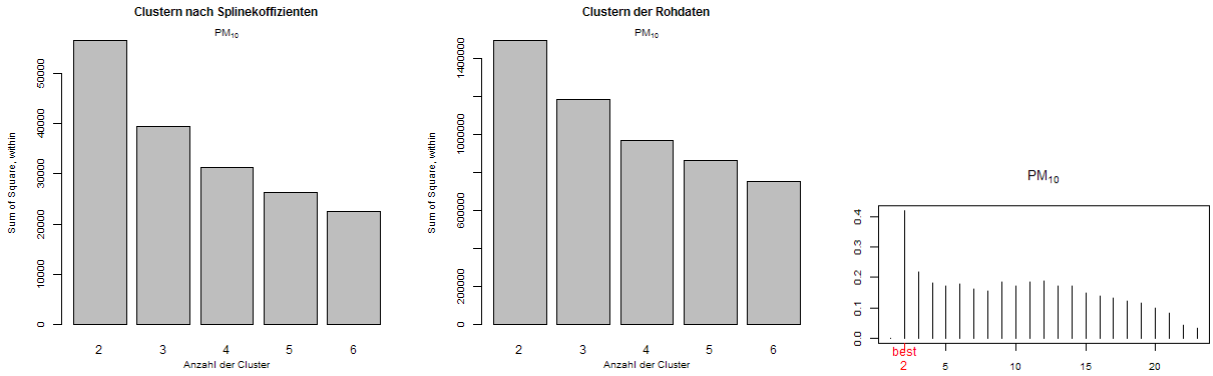


Abb. B.2.1. PM_{10}: Screeplots und Plot der durchschnittlichen Silhouette-Breiten

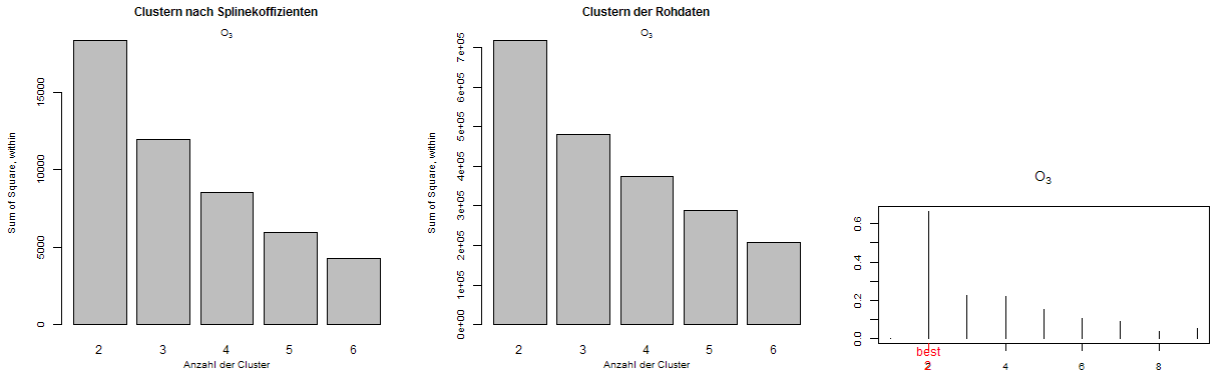


Abb. B.2.2. O_3: Screeplots und Plot der durchschnittlichen Silhouette-Breiten

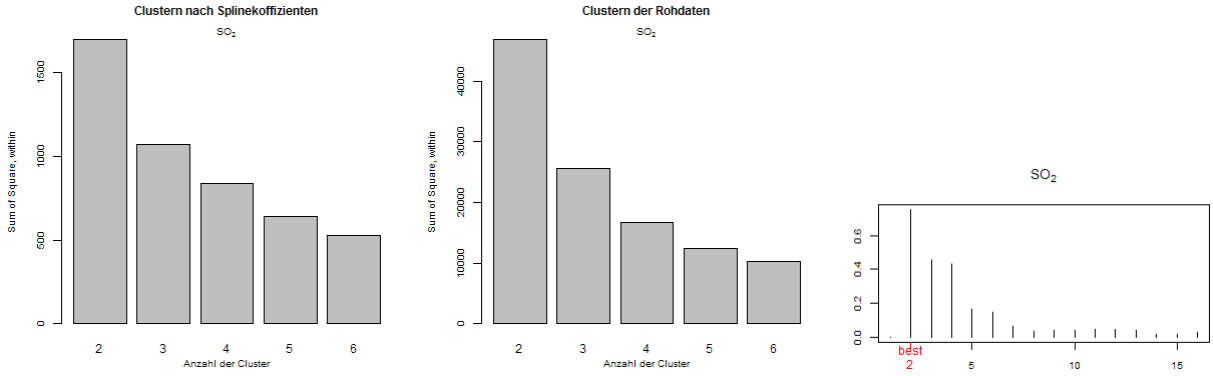


Abb. B.2.3. SO_2: Screeplots und Plot der durchschnittlichen Silhouette-Breiten

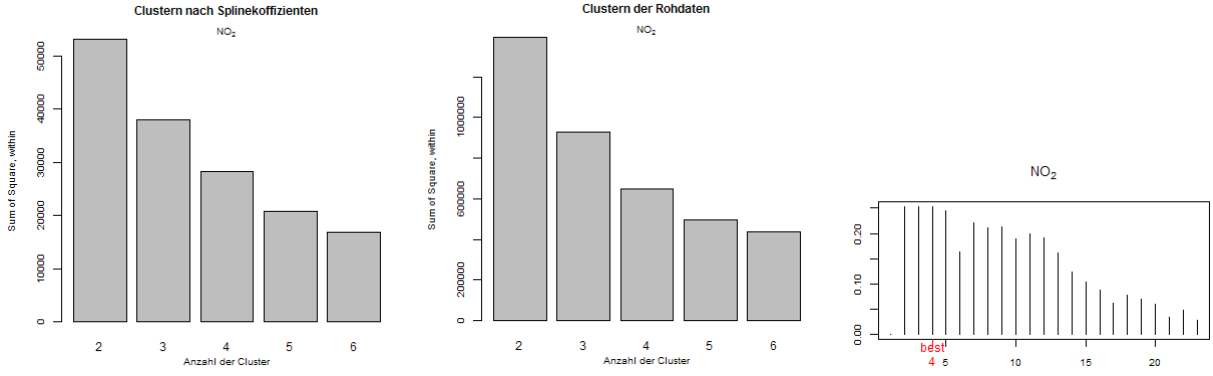


Abb. B.2.4. NO_2: Screeplots und Plot der durchschnittlichen Silhouette-Breiten

B.3 Silhouette-Plots für verschiedene Klassenanzahl

B.3.1 Feinstaub PM_{10} [µg/m³]

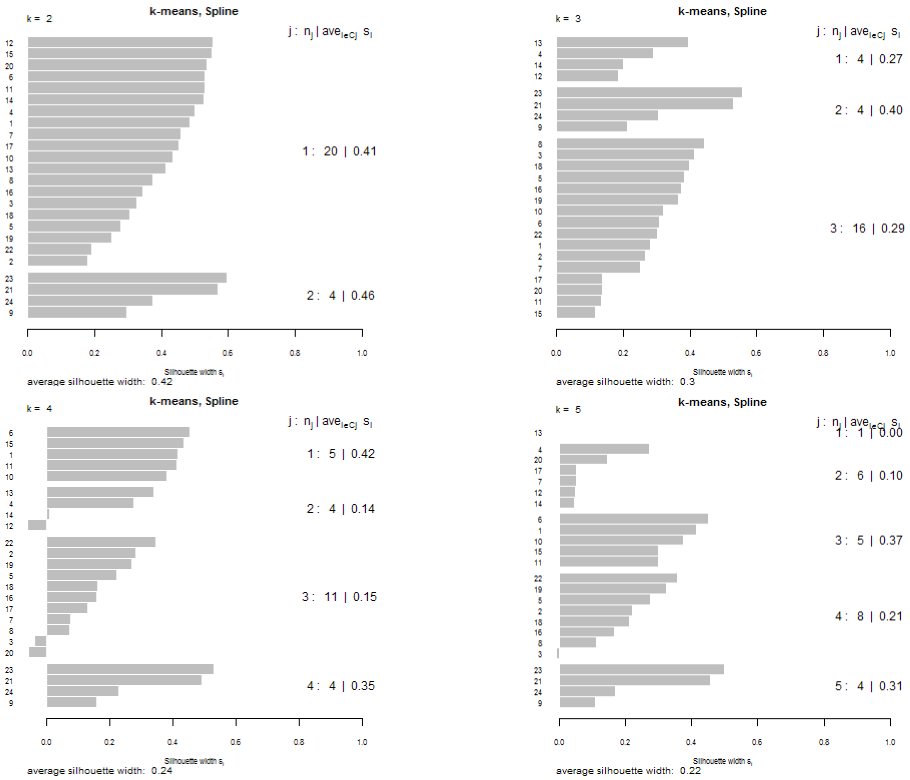


Abb. B.3.1. PM_{10}: Silhouette für *k* = 2, …, *5*, *k-means* der Splinekoeffizienten

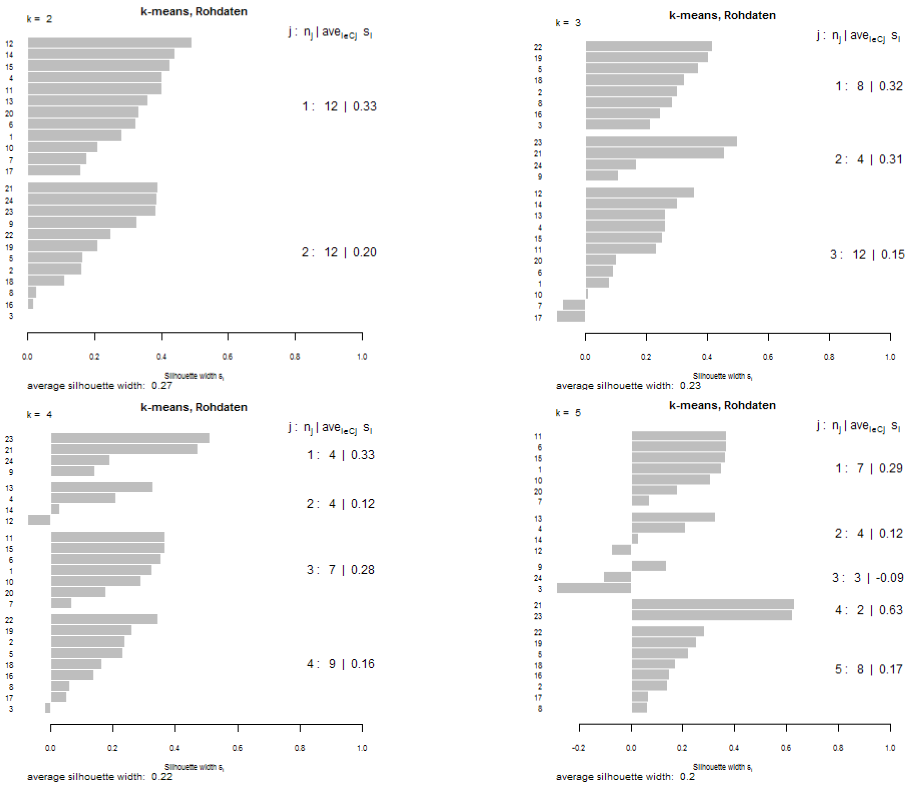


Abb. B.3.2. PM_{10}: Silhouette für *k* = 2, …, *5*, *k-means* der Rohdaten

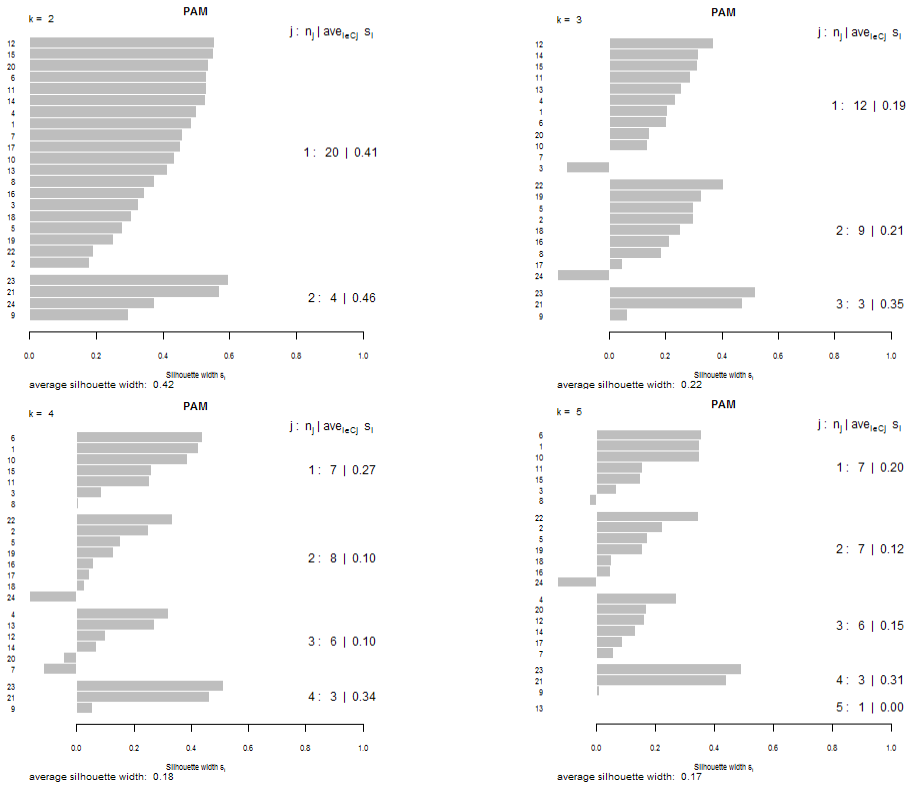


Abb. B.3.3. PM_{10}: Silhouette für *k* = 2, …, 5, *PAM*

B.3.2 Ozon O_3 [µg/m³]

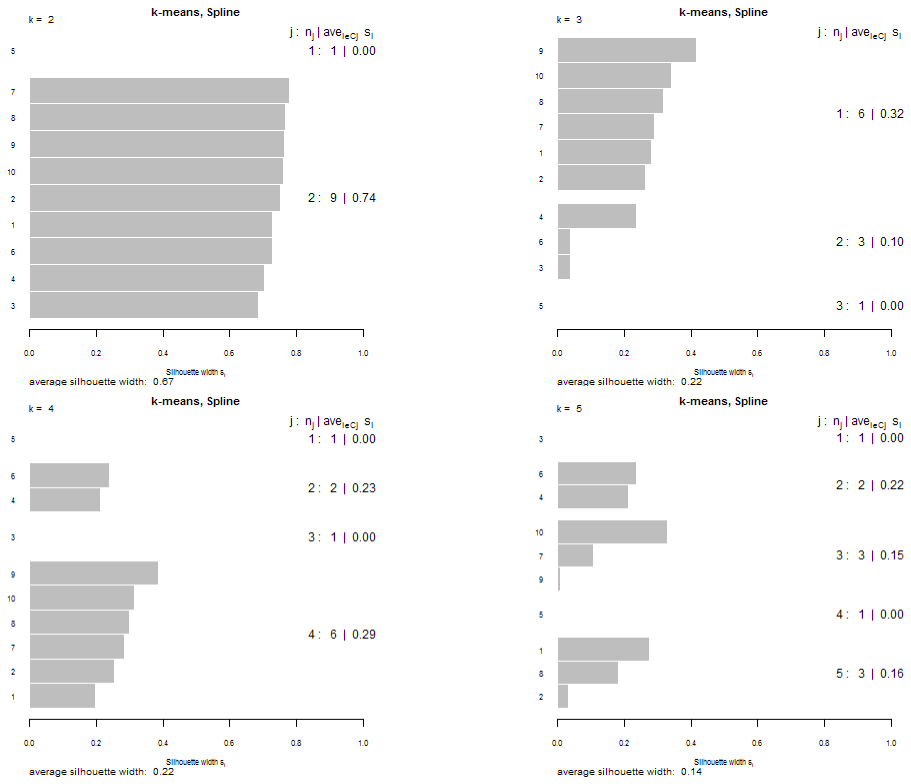


Abb. B.3.4. O_3: Silhouette für *k* = 2, …, 5, *k-means* der Splinekoeffizienten

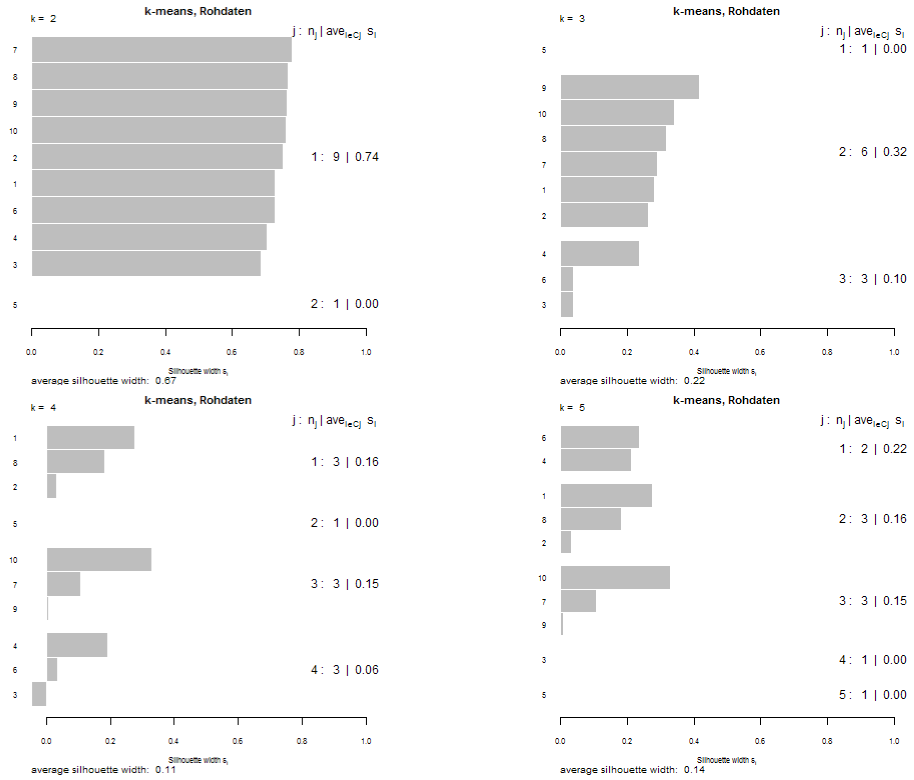


Abb. B.3.5. O_3: Silhouette für *k* = 2, …, 5, *k-means* der Rohdaten

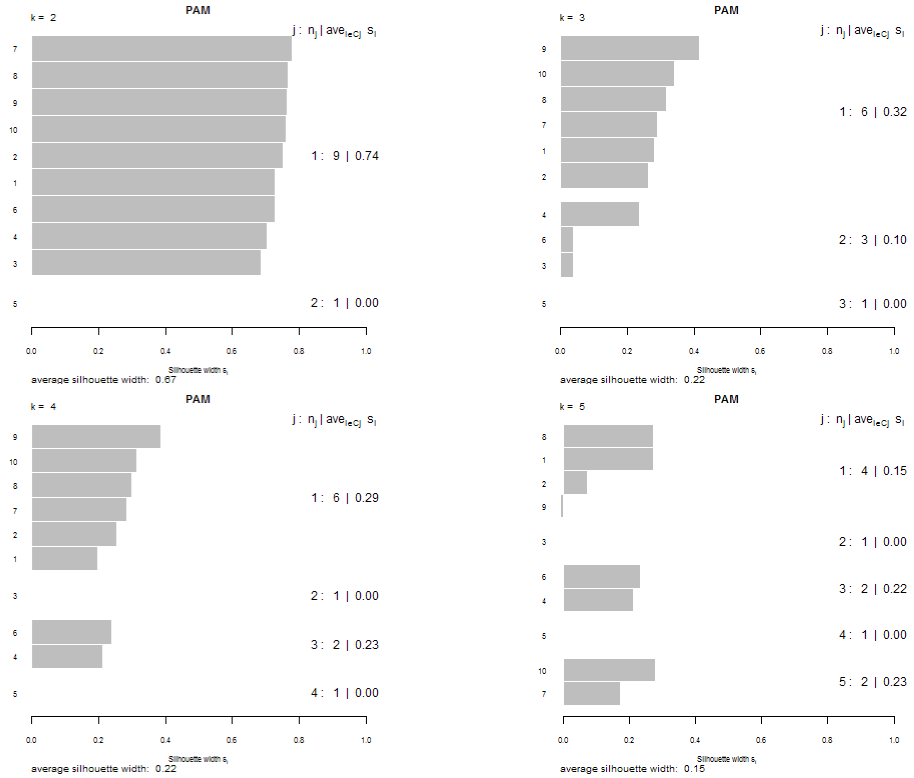


Abb. B.3.6. O_3: Silhouette für *k* = 2, …, 5, *PAM*

B.3.3 Schwefeldioxid SO_2 [µg/m³]

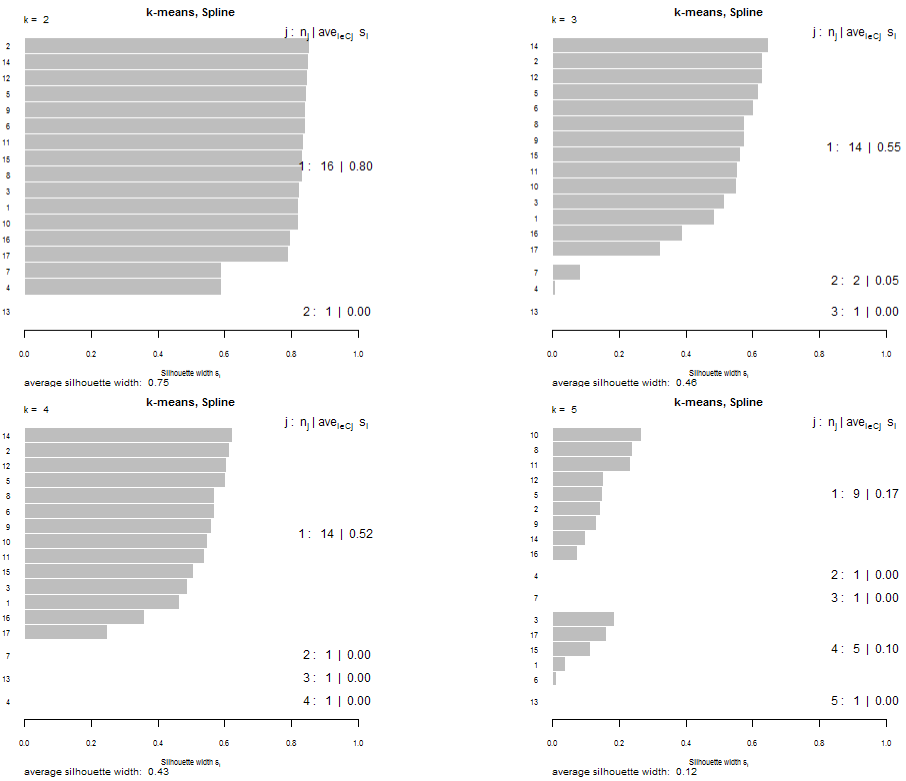


Abb. B.3.7. SO_2: Silhouette für *k* = 2, …, 5, *k-means* der Splinekoeffizienten

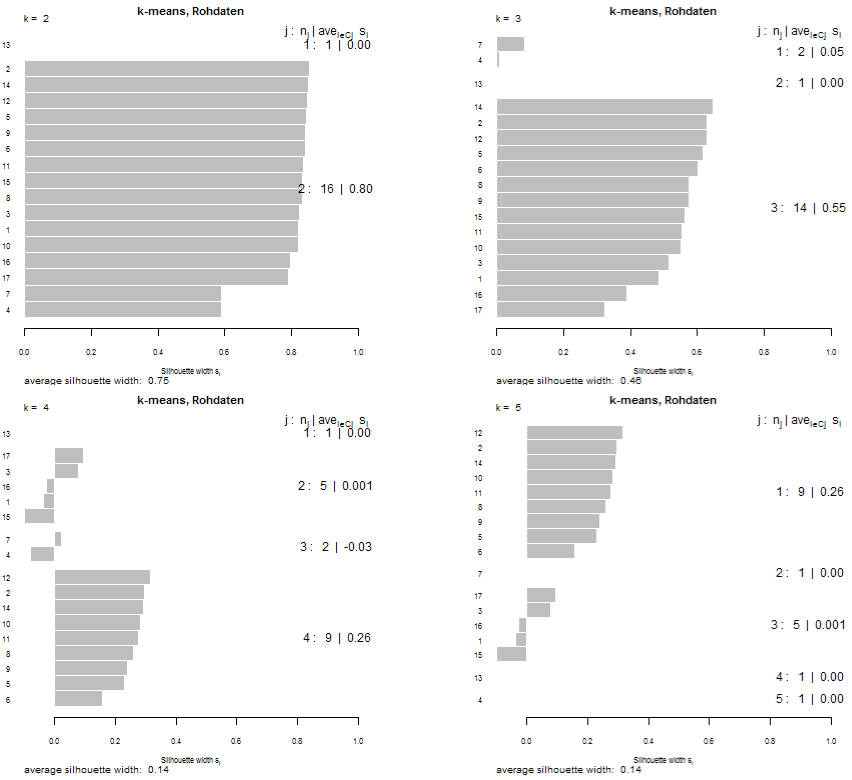


Abb. B.3.8. SO_2: Silhouette für *k* = 2, …, 5, *k-means* der Rohdaten

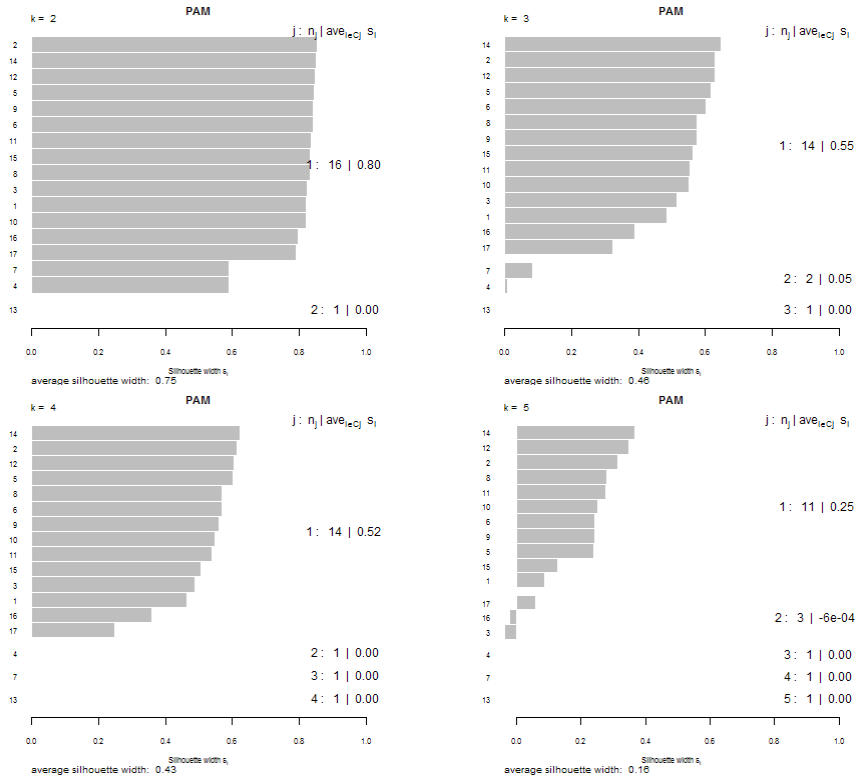


Abb. B.3.9. SO_2: Silhouette für *k* = 2, …, 5, *PAM*

B.3.4 Stickstoffdioxid NO_2 [µg/m³]

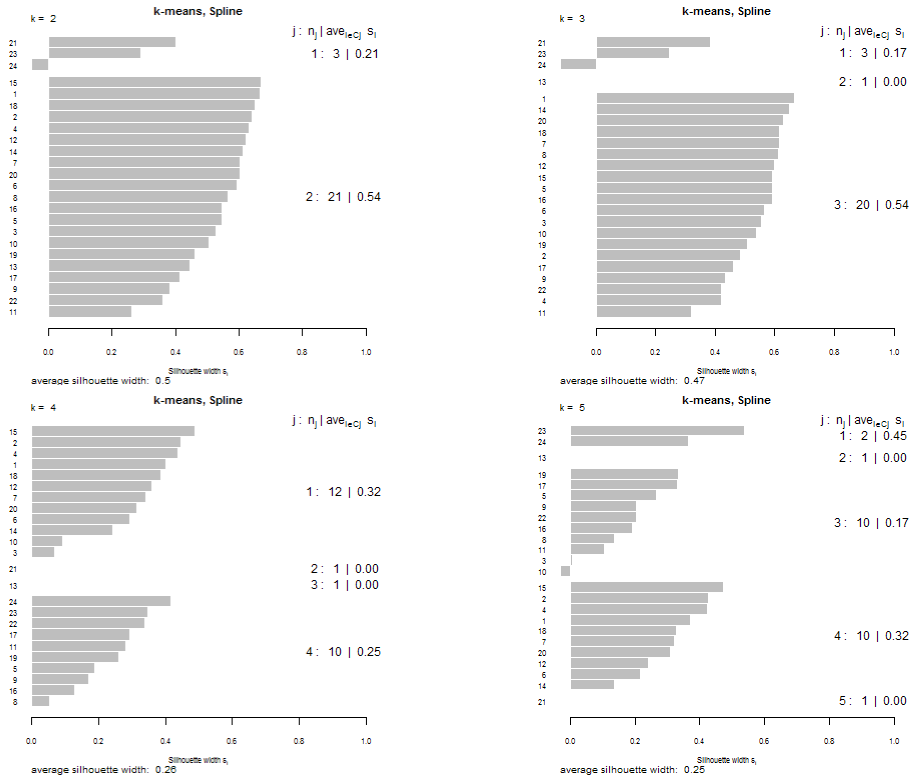


Abb. B.3.10. NO_2: Silhouette für *k* = 2, …, 5, *k-means* der Splinekoeffizienten

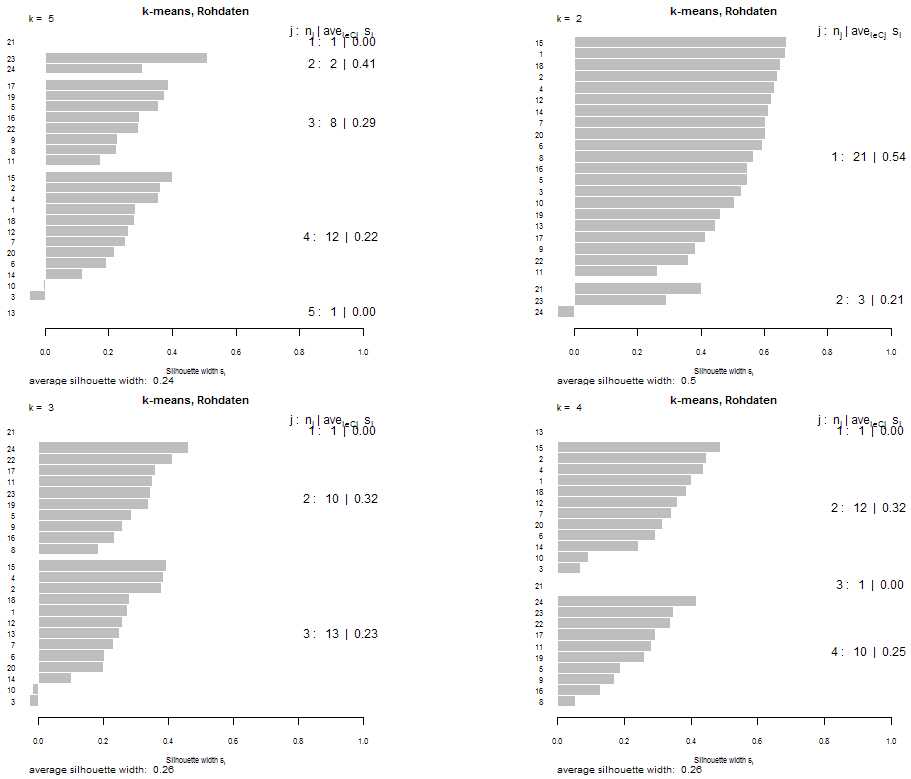


Abb. B.3.11. NO_2: Silhouette für *k* = 2, …, *5*, *k-means* der Rohdaten

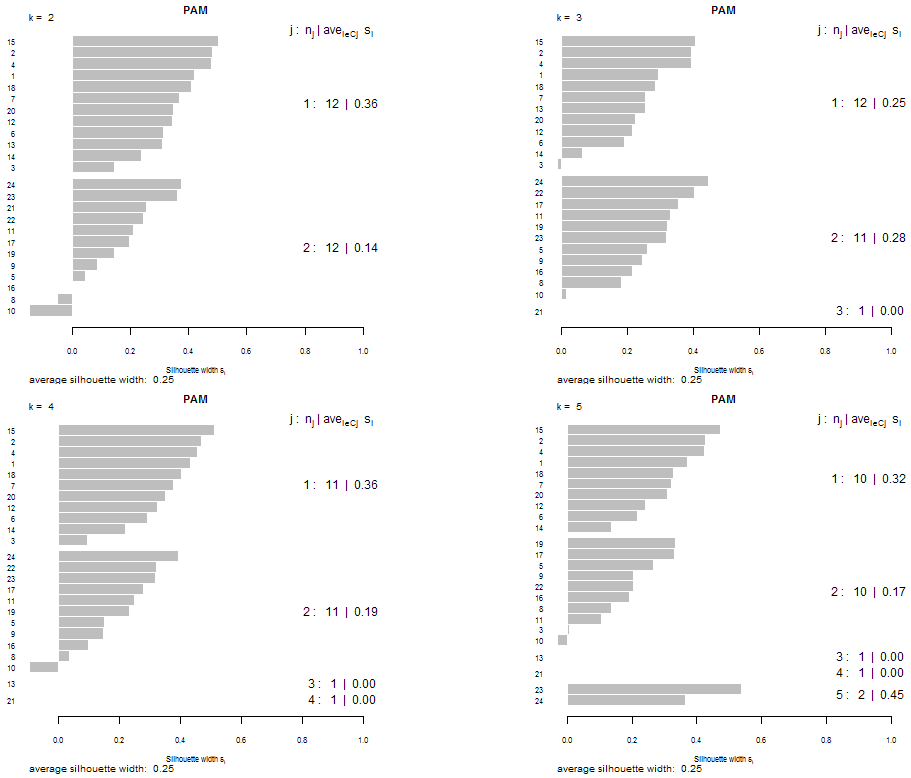


Abb. B.3.12. NO_2: Silhouette für *k* = 2, …, *5*, *PAM*

C Umsetzung mit der Statistik Software *R*

Die Messwerte liegen in Form einer csv-Datei (TMW08_10.csv) mit 12 Spalten vor. Spalten 2 – 5 enthalten die Beobachtungen zu den Luftschadstoffen PM_{10} [µg/m³], O_3 [µg/m³], SO_2 [µg/m³] und NO_2 [µg/m³] und in Spalte 12 ist die Messstation vermerkt (siehe Tabelle C.1)

Spalte	*Merkmal*	*Typ*	*Codierung*
1	Nr.		
2	**PM_{10}**	**stetig**	
3	**O_3**	**stetig**	
4	**SO_2**	**stetig**	
5	**NO_2**	**stetig**	
6	Tag	kategorisch	1 – 31 (Kalendertag)
7	Monat	kategorisch	1 – 12 (Jän. – Dez.)
8	Jahr	kategorisch	2008, 2009, 2010 (Jahr der Messung)
9	Wochentag	kategorisch	1… Mo-Fr, 2… Sa, 3… So
10	Feiertag	kategorisch	1… Arbeitstag, 2… Sa, 3…So oder Feiertag
11	So_Wi	kategorisch	1… Sommer (April – Sept.), 2… Winter (Okt. – März)
12	**Station**	**kategorisch**	**1 – 25**

Tab. C.1: Merkmale und Codierung der Daten der csv-Datei (TMW08_10.csv)

C.1 Einlesen und Aufbereiten der Daten

```
setwd("C:/")
luft<-read.csv("TMW08_10.csv",header=TRUE,sep=";",fill=TRUE,comment.char="")

station <- c("Bruck","Deutschlandsberg","Fürstenfeld","Judenburg","Judendorf", "Kapfenberg",
  "Knittelfeld","Köflach","Leibnitz","Leoben/Donawitz","Leoben/Göß","Liezen","Masenberg",
  "Mürzzuschlag","Niklasdorf","Peggau","Straßengel","Voitsberg","Weiz","Zeltweg","Graz Don
   Bosco","Graz Mitte Gries","Graz Nord",  "Graz Süd","Graz West")

# Für die Bezeichnung der Stationen müssen bei O3 und SO2 Stationen ausgeschlossen werden
station_O3 <- c(2,3,4,12,13,14,18,19,23,24)
station_SO2 <- c(1,2,3,5,6,7,8,9,10,11,12,13,15,16,17,18,23,24,25)
```

function reading: Einlesen der Daten
Eingabe: L Luftschadstoff
Ausgabe: m Zeile :Tag 1 … 1096 Spalte: Station 1 … 25

```
reading <- function(L) {
  m <- matrix(nrow=1096, ncol=25); colnames(m) <- station
  for (i in 1:25)  { m[,i] <- luft[luft[,12]==i, L] }
  return(m) }
```

function missingvalues: Ersetzen der fehlenden Werte durch Stationsmittelwert

```
missingvalues <- function(daten, n) {
  for (i in 1:n)  { daten[is.na(daten[,i]), i] <- mean(na.omit(daten[,i])) }
  return(daten) }
```

```
PM10_mv <- reading(2);O3_mv <- reading(3); SO2_mv <- reading(4); NO2_mv <- reading(5);

# Entfernen der Stationen, die keine Daten liefern
O3_mv <- O3_mv[,c(-1,-5,-6,-7,-8,-9,-10,-11,-15,-16,-17,-20,-21,-22,-25)]    ,
SO2_mv <- SO2_mv[,c(-4,-14,-19,-20,-21,-22)]

PM10 <- missingvalues(PM10_mv,25);  O3 <- missingvalues(O3_mv,10);
SO2 <- missingvalues(SO2_mv,19);       NO2 <- missingvalues(NO2_mv,25)
```

C. 2 Approximation der Funktionen x_i

Das Package *fda* der Statistik Software *R* bietet die Möglichkeit, Funktionen mithilfe der Routine smooth.basis(argvals,y,fdParobj) zu glätten und die entsprechenden Basisfunktionen mit create.bspline.basis(rangeval,nbasis,norder,breaks) zu erzeugen (vgl. Ramsay/Hooker/Graves 2009, S. 29-35 bzw. S. 46-54). Die Knoten können dabei entweder direkt eingegeben werden (über breaks) oder sie werden mithilfe der Anzahl der Basisfunktionen (= nbasis) ermittelt.

C.2.1 Erzeugung der Basisfunktionen

```
library(fda)

knoten <- c(1,32,61,92,122,153,183,214,245,275,306,336, 367,398,426,457,487,518,548,579,610,
640,671,701, 732,760,791,822,852,883,913,944,975,1005,1036,1066,1096)

Basis16 <- create.bspline.basis(c(1,1096), nbasis=3*4+4)        # Knoten = 1/Quartal
Basis22 <- create.bspline.basis(c(1,1096), nbasis=3*6+4)        # Knoten = alle 2 Monate
Basis40 <- create.bspline.basis(c(1,1096), nbasis=3*12+4)       # Knoten = 1/Monat
Basis76 <- create.bspline.basis(c(1,1096), nbasis=6*12+4)       # Knoten = 2/Monat
KBasis <- create.bspline.basis(norder=4, breaks=knoten)         # Knoten am Monatsersten
```

C.2.2 Glättung der Funktionen

```
zeitachse <- matrix(nrow=1096, ncol=1, 1:1096)

PM10_16fd <- smooth.basis(zeitachse, PM10, Basis16)$fd
PM10_22fd <- smooth.basis(zeitachse, PM10, Basis22)$fd
PM10_40fd <- smooth.basis(zeitachse, PM10, Basis40)$fd
PM10_76fd <- smooth.basis(zeitachse, PM10, Basis76)$fd
PM10_Kfd  <- smooth.basis(zeitachse, PM10, KBasis)$fd
```

C.2.3 Graphische Darstellung der geschätzten Funktionen

Plotten der geschätzten Funktionen für den Luftschadstoff PM_{10} bei Verwendung von unterschiedlichen Basen

function plotfunction: Plotten von Funktionen

```
plotfunction <- function(daten, basisknoten, varname, untertitel, ygrenze) {
  par(mar=c(2,4,3,1))
```

```
  plot(daten,xaxt="n",xlab="",lty=1,col=1,cex.axis=0.75,cex.lab=0.8, ylab=varname, ylim=ygrenze)
  mtext(expression(paste("  ","["mu*g/m^3,"]")), side=2, line=3, adj=0.8, cex=0.8)
  lines(mean(daten), col="red", lwd=2)
  abline(v = basisknoten, col="gray", lty="dotted")
  abline(v = c(1,1096), col="gray", lty="dotted")
  title(main="Umwandlung mithilfe Regressions-Spline", cex.main=0.9)
  mtext(untertitel, side=3, cex=0.8)
  axis(1,zeitknoten, labels=zeitlabel,cex.axis=0.7, lwd=1, col.ticks="gray")   }

plotfunction(PM10_16fd,Basis16$params, expression(PM[10]), "äqidistante Knoten, vierteljährlich" ,
    c(0,150))
```

Plotten der geschätzten Funktion mit 36 inneren Knoten einer Messstation für den Luftschadstoff PM_{10}

```
png(filename="PM_Approximation%03d.png", height=260, width=340, restoreConsole=T)
for (i in 1:25) {
  plot(PM10_40fd[i], xaxt="n", ylab=expression(PM[10]), xlab="", col="red", lwd=2, cex.axis=0.7,
        cex.lab=0.75, ylim=c(0,100))
  title(main=station[i], cex.main=1)
  mtext(expression(paste("  ","["mu/m^3,"]")), side=2, line=3, adj=0.8, cex=0.7)
  axis(1, zeitknoten2, labels=zeitlabel2, cex.axis=0.7, lwd=1, col.ticks="gray")
  points(zeitachse, PM10_mv[,i], type="p", cex=0.4) }
dev.off()
```

C.3 Clusterung

```
library(cluster)
library(flexclust)

# Fürs Clustern wird die Station Graz Mitte entfernt
PM10 <- PM10[,-22]; NO2 <- NO2[,-22]
station2 <- c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,23,24,25)
colnames(PM10) <- station[station2]; colnames(NO2) <- station[station2]
```

C.3.1 Bestimmung der Koeffizientenmatrix

```
# Werte der Basisfunktionen
Basismat <- eval.basis(as.vector(zeitachse), Basis40)
Basismat22 <- eval.basis(as.vector(zeitachse), Basis22)
Basismat76 <- eval.basis(as.vector(zeitachse), Basis76)
```

 # geschätzte Koeffizienten (Kleinste-Quadrate-Schätzer): $\hat{c} = \left(\Phi^T\Phi\right)^{-1}\Phi^T y$

```
# y n-dim. Vektor der Beobachtungen,
# Φ n x k Matrix  mit den Basisfunktionswerte Φk(tj)

PM10coeff <- solve(crossprod(Basismat), crossprod(Basismat, PM10))
PM10co22 <- solve(crossprod(Basismat22), crossprod(Basismat22, PM10))
PM10co76 <- solve(crossprod(Basismat76), crossprod(Basismat76, PM10))

# Erzeugung der Funktionen xi mithilfe der Basisfunktionen der geschätzten Koeffizienten
PM10fd <- fd(PM10coeff, Basis40)
PM10fd22 <- fd(PM10co22, Basis22); PM10fd76 <- fd(PM10co76, Basis76)
```

C.3.2 Clusterung der Splinekoeffizienten mit *k-means*

```
PSp2<-kmeans(t(PM10fd$coef), 2, nstart=20); PSp3<-kmeans(t(PM10fd$coef), 3, nstart=20);
PSp4<-kmeans(t(PM10fd$coef), 4, nstart=20); PSp5<-kmeans(t(PM10fd$coef), 5, nstart=20);
PSp22<-kmeans(t(PM10fd22$coef), 3, nstart=20); PSp76<-kmeans(t(PM10fd76$coef), 3, nstart=20)

# Erzeugung der (geglätteten) Funktionen der Clusterzentren
PSp2fd <- fd(t(PSp2$centers), Basis40);   PSp3fd <- fd(t(PSp3$centers), Basis40)
PSp4fd <- fd(t(PSp4$centers), Basis40);   PSp5fd <- fd(t(PSp5$centers), Basis40)

PM10_Spline <- cbind(PSp2$cluster, PSp3$cluster, PSp4$cluster)
write.csv(PM10_Spline, file="PM10_kmSpl.csv")
```

C.3.3 Clusterung der Rohdaten mit *k-means*

```
Pk2 <- kmeans(t(PM10), 2, nstart = 20);   Pk3 <- kmeans(t(PM10), 3, nstart = 20)
Pk4 <- kmeans(t(PM10), 4, nstart = 20);   Pk5 <- kmeans(t(PM10), 5, nstart = 20)
Pk22 <- kmeans(t(PM10), 3, nstart = 20); Pk76 <- kmeans(t(PM10), 3, nstart = 20)

# Erzeugung der (geglätteten) Funktionen der Clusterzentren
Pk2fd <- smooth.basis(zeitachse, t(Pk2$centers), Basis40)$fd
Pk3fd <- smooth.basis(zeitachse, t(Pk3$centers), Basis40)$fd
Pk4fd <- smooth.basis(zeitachse, t(Pk4$centers), Basis40)$fd
Pk5fd <- smooth.basis(zeitachse, t(Pk5$centers), Basis40)$fd

PM10_Roh <- cbind(Pk2$cluster, Pk3$cluster, Pk4$cluster)
write.csv(PM10_Roh, file="PM10_kmRoh.csv")
```

C.3.4 Clusterung der Splinekoeffizienten mit *PAM*

```
PP2 <- pam(t(PM10fd$coef), 2);   PP3 <- pam(t(PM10fd$coef), 3)
PP4 <- pam(t(PM10fd$coef), 4);    PP5 <- pam(t(PM10fd$coef), 5)
PP22 <- pam(t(PM10fd22$coef), 3);   PP76 <- pam(t(PM10fd76$coef), 3)

# Erzeugung der (geglätteten) Funktionen der Clusterzentren
PP2fd <- fd(t(PP2$medoids), Basis40);  PP3fd <- fd(t(PP3$medoids), Basis40)
PP4fd <- fd(t(PP4$medoids), Basis40);  PP5fd <- fd(t(PP5$medoids), Basis40)

PM10_PAM <- cbind(PP2$cluster, PP3$cluster, PP4$cluster)
write.csv(PM10_PAM, file="PM10_pam.csv")
```

C.3.5 Berechnung des RandIndex

function randTable: Berechnung der RandIndices für 3 Clusterungen und Speichern in Tabelle
Eingabe: clus1,…,clus3   Clusterung
namen Bezeichnung der Methoden (für Zeilen bzw. Spaltenbezeichnung)

```
randTable <- function(clus1, clus2, clus3, namen)  {
  Tab <- matrix(nrow=2, ncol=2, dimnames=namen)
  Tab[1,1] <- adjustedRandIndex(clus1, clus2)
  Tab[1,2] <- adjustedRandIndex(clus1, clus3)
  Tab[2,2] <- adjustedRandIndex(clus2, clus3)
  return(Tab) }

# Vergleich: k-means der Splinekoeffizienten, k-means der Rohdaten, PAM und k = 3,4
methode1 <- c("k-means 2", "PAM");  methode2 <- c("k-means 1", "k-means 2")
```

```
RandPM10_3 <- randTable(PSp3$cluster, Pk3$cluster, PP3$cluster, list(methode2,methode1))
RandPM10_4 <- randTable(PSp4$cluster, Pk4$cluster, PP4$cluster, list(methode2,methode1))
RandPM10 <- cbind(RandPM10_3, RandPM10_4)

write.csv(round(RandPM10,2), file="PM10_aRand.csv")

# Vergleich: Clusterungmethoden  für unterschiedliche Basen
BasisName <- list(c("Basis40", "Basis22"), c("Basis22", "Basis76"))
RandPM10_k1 <- randTable(PSp3$cluster, PSp22$cluster, PSp76$cluster, BasisName)
RandPM10_k2 <- randTable(Pk3$cluster, Pk22$cluster, Pk76$cluster, BasisName)
RandPM10_pam <- randTable(PP3$cluster, PP22$cluster, PP76$cluster, BasisName)
RandPM10B <- cbind(RandPM10_k1, RandPM10_k2, RandPM10_pam)

write.csv(round(RandPM10B,2), file="PM10Basis_aRand.csv")
```

C.3.6 Graphische Darstellung der Screeplots bzw. Silhouette-Plots

function scree.plot: Plotten des Screeplot für verschiedene Klassenanzahl k

Eingabe: *k1,…,k5* *Clusterung (mit k-means)*
methode *k-means der Splinekoeffizienten oder k-means der Rohdaten*
varname[26] *Bezeichnung des Luftschadstoffes*

```
scree.plot <- function(k1, k2, k3, k4, k5, methode, varname)  {
  wss <- c(sum(k1$withinss), sum(k2$withinss),sum(k3$withinss),sum(k4$withinss), sum(k5$withinss))
  names(wss) <- 2:6
  barplot(wss, cex.axis=0.8, cex.lab=0.8, ylab="Sum of Square, within")
  title(Methode, cex.main=1)
  mtext(varname, side=3, line=0, adj=0.5, cex=0.8)
  mtext("Anzahl der Cluster", side=1, line=2, cex=0.8) }

scree.plot(PSp2, PSp3, PSp4, PSp5, PSp6, "Clustern nach Splinekoffizienten" , expression(PM[10]))
```

function plot.silhouette.best: Plotten der durchschnittlichen Silhouette-Breiten für verschiedene Klassenanzahl k

Eingabe: *clusdaten* *Daten für Clusterung*
maxk *maximale Klassenanzahl der Clusterung*

```
plot.silhouette.best <- function(clusdaten, maxk, varname)  {
  asw <- numeric(n)
  for (k in 2:maxk) { asw[k] <- pam(clusdaten, k) $ silinfo $ avg.width }
  k.best <- which.max(asw)
  cat("silhouette-optimal number of clusters:", k.best, "\n")
  plot(1:maxk, asw, type= "h", main = varname, xlab= "k  (# clusters)", cex.lab=0.8,
     ylab = "average silhouette width", cex.axis=0.8)
  axis(1, k.best, paste("best", k.best, sep="\n"), col="red", col.axis="red", cex=0.8) }

plot.silhouette.best(t(PM10fd$coef), 23, expression(PM[10]))
```

[26] Variablen mit derselben Bedeutung werden in allen *functions* gleich bezeichnet und in weiterer Folge nicht mehr extra genannt.

function plot.sil: Plotten der Silhouette-Breiten für jedes Objekt der Clusterung

Eingabe: *clus* *Clusterung*
distmat *Distanzmatrix*
k *Klassenanzahl*
titel *Titel der Graphik*

```
plot.sil <- function(clus, distmat, k, titel)  {
   sil <- silhouette(clus$cluster, distmat)
   plot(sil, cex.lab=0.7, cex.axis=0.7, cex=0.7, main="",do.n.k=F)
   title(main=titel, cex.main=0.9)
   mtext(paste("k = ",k), side=3, adj=0, cex=0.8)
   mtext(paste("average silhouette width:", round(summary(sil)$avg.width,2)),side=1, adj=0,
   line=3, cex=0.8) }

PMdist <- dist(t(PM10fd$coef), "euclidean")
plot.sil(PSp2fd, PMdist, 2, "k-means, Splines")
```

C.3.7 Graphische Darstellung der Cluster

function plot.cluster: Plotten eines Clusters mit den geglätteten Funktionen der Messstationen, die sich in diesem Cluster befinden

Eingabe: *clusfd* *Clusterung als funktionale Daten (Typ: fd)*
daten *Beobachtungen*
cln *Nummer des Clusters*
n *Anzahl der Messstationen*
ygrenze *max. Wert für die Beobachtungen*
mitte *für horizontale Linie*
m *Position von „varname"*

```
plot.cluster <- function(clustfd, clust, daten, cln, n, titel, varname, ygrenze, mitte, m) {
  plot(clustfd[cln], col=2, lwd=2, xaxt="n", xlab="", ylab=varname, ylim=ygrenze)
  title(main=titel, cex.main=1.5)
  mtext(varname, side=3, line=0.9, adj=m, cex=1)
  abline(h=mitte, lty="dotted")
  axis(1, zeitknoten,  labels=zeitlabel, cex.axis=0.7,  lwd=1, col.ticks="gray")
  mtext(paste("Cluster ", cln), line=-1.2, adj=1, cex=0.75)
  for (i in 1:n) { if(clust$cluster[i]==cln) { lines(daten[i]) }  } }

# Darstellung der 3 Cluster in einer Graphik
for (i in 1:3) { plot.cluster(PSp3fd, PSp3, PM10fd, i, 24, "k-means (Splinekoeffizienten) - ",
    expression(PM[10]), c(0,80), 40, 0.71) }
```

function plot.zentren: Plotten der Clusterzentren mit den geglätteten Funktionen

Eingabe: *untertitel* *Ergänzung zum Titel*
(Rest wie bei den anderen functions)

```
plot.zentren <- function(clustfd, daten,  varname, untertitel, ygrenze, mitte) {
  par(mar=c(2,4,3,2))
  plot(daten, xaxt="n", xlab="", ylab=varname,l ty=1, col="azure4", cex.axis=0.7, cex.lab=0.7,
       ylim=ygrenze)
  title(main="Clusterzentren und funktionale Daten", cex.main=1)
  mtext(untertitel, side=3, cex=0.8)
  abline(h=mitte, lty="dotted")
  axis(1, zeitknoten, labels=zeitlabel, cex.axis=0.7, lwd=1, col.ticks="gray")
  lines(clustfd, lwd=2, col=2, lty=1) }
plot.zentren(PSp3fd, PM10fd, expression(PM[10]), "k-means (Splinekoeffizienten) mit k =
3", c(0,80), 40)
```